IBM PowerVM 实施手册

文档版本: V0.1

文档日期:

文档修订信息

文档名称	IBM PowerVM 实施手册
作者	

修订历史 (REVISION HISTORY)				
版本	修 正 章 节	日期	作者	变更纪录

文档核准信息

版本	核准日期	核准人	所属单位部门	备注

目 录

1 前言

1.1 编写目的

本文旨在提供一个 PowerVM 操作的指导手册，而不是 PowerVM 的基础介绍，假设读者有一定的 PowerVM 基本知识，具体包含如下：

- PowerVM 的规划
- PowerVM 的安装配置
- PowerVM 的管理
- PowerVM 中网络和存储的配置
- PowerVM 环境下性能的监控
- PowerVM 的一些高级特性介绍

本文仅针对 HMC 环境下的 PowerVM 配置，不包括 IVM 的配置，且客户端分区为 AIX，不包含 Power Linux 和 i 系统的安装配置，

随着技术的不断更新，本文也会继续完善更新。

1.2 适用范围

PowerVM 实施者，维护者。

2 IBM PowerVM 介绍

2.1 PowerVM 基本特性

PowerVM 是一个包含一系列硬件和软件特性的品牌，它使系统能够更灵活地适应各种工作负载。它包含微分区、逻辑分区、虚拟 I/O、微码技术、活动分区迁移、动态内存扩展等众多特性。这些新特性几乎可以将服务器上所有的物理资源虚拟化，并为客户提供更好的高可用性和资源利用率。目前 PowerVM 有三个版本，分别是易捷版 PowerVM Express Edition，标准版 PowerVM Standard Edition，和企业版 PowerVM Enterprise Edition。

PowerVM Editions	**Express**	**Standard**	**Enterprise**
Concurrent VMs	**3 per server**	**20 per core (up to 1000)**	**20per core (up to 1000)**
Management	**IVM**	**IVM,HMC**	**IVM,HMC**
Virtual I/O Server	✓	✓ ✓	✓ ✓
NPIV	✓	✓	✓
Suspend/Resume		✓	✓
Multiple Shared Processor Pools		✓	✓
Thin Provisioning		✓	✓
Live Partition Mobility			✓
Active Memory Sharing			✓
Shared Storage Pools		✓	✓
Active Memory Deduplication			✓
Network Balancing		✓	✓
Live Partition Mobility Performance Improvements			✓

2.1.1 PowerVM 术语介绍

下面就 PowerVM 的主要组成部分做一个基本的介绍：

PowerVM Hypervisor: PowerVM Hypervisor 是 PoweVM 的基础，是 PowerVM 的系统管理程序，控制资源的分配，是 Power 服务器底层提供的功能，在一个系统上支持多个操作环境。

Micro-Partitioning：微分区也是将一个物理的服务器划分成若干个服务器,一个微分区可以以 1/10 个(最新的 Power7+ CPU 支持最小颗粒度为 1/20)物理处理器来进行分配,它允许多个分区共享一组物理处理器的处理资源。

Dynamic Logical Partitioning:动态逻辑分区(DLPAR)技术是 IBM Power 服务器的一项重要的虚拟化特性，可以在分区运行的情况下动态调整逻辑分区资源的分配，从而满足不断变化的业务需求。

Shared Processor Pools: 共享处理器池是先将物理 CPU 放置于一个共享的处理器池中并统一的被分配到微分区之中使用，现在的 Power 服务器可以支持多个共享处理器池。

Shared Storage Pools： 共享存储池是一个 SAN 存储设备的池，这些设备可以跨多个 VIO 服务器，其中多个 VIO 服务器组成一个集群，使用共享的 SAN 存储设备，当 VIO 服务器启用共享存储池的时候，VIO 服务器通过逻辑单元（logical units）来使用存储池中的空间,并将逻辑单元通过 vSCSI 映射给客户端分区使用。可以将 Power Systems 服务器和 VIOS 的存储资源集中到池中，以便优化资源利用率。

Live Partition Mobility： 实时分区迁移，可以在服务器之间移动实时 AIX、Linux 和 IBM i 虚拟机，从而消除计划内的停机。

Active Memory Sharing：共享内存是分配给共享内存池的物理内存，在多个逻辑分区之间共享该内存。共享内存池是已定义的一组物理内存块，它们由系统管理程序作为单个内存池进行管理。配置

为使用共享内存的逻辑分区（以后称为共享内存分区）与其他共享内存分区共享池中的内存。

Active Memory Deduplication：AMD(Active Memory Deduplication)是 PowerVM 中的一项新技术。其基于 AMS（Active Memory Sharing)，通过对在 AMS 内存共享池内存储内容相同的内存空间进行去重，来达到优化内存使用的目的。Hypervisor 对在 AMS 共享池内的内存页进行比对，如果发现相同内容的内存页，则通过映射来释放重复的内存空间，提高物理内存的使用效率。AMD 支持 AIX，i 和 Linux 分区。其对硬件的微码要求为 740_042, 目前只有最新的 POWER7 C 类 Server 使用该级别的微码。

NPIV：N_Port ID Virtualization (NPIV) 是业界通行的一个工业标准，主要目的是通过虚拟化光纤通道接口来简化 SAN 网络的架构 ，让虚拟环境下的服务器与 SAN 环境连接更加弹性和安全。NPIV 即为光纤信道中的一个协议，目的在于让一个实体的 N 端口可以虚拟出数个 N_Port ID。并且将光纤交换机上的任一 F_Port 关联到多个 N_Port ID, 从而使基于 Power 虚拟化平台的多个不同的分区系统可以共享一个光纤适配卡 (HBA)。其功能是由光纤信道 HBA 卡提供，但前端的虚拟平台以及后端的光纤信道交换机也要能支持。简化管理并提高光纤通道 SAN 环境的性能。

VIOS：PowerVM 虚拟化功能部件的关键组件之一是 VIOS(Virtual IO Server)虚拟 I/O 服务器，是一个特殊定制化的 AIX 操作系统的逻辑分区，可以提供存储 IO 和网络资源的虚拟化，从而使得各个逻辑分区通过 VIOS 来共享这些物理资源。

Virtual Ethernet：虚拟以太网并不需要 PowerVM。它让两个 LPAR 可以通过 Hypervisor 和 Vrtual Ethernet 通道通信。Virtual Ethernet 在传输网络通信流时需要开销一些 CPU 和内存资源。它还支持 Virtual LAN 和其他安全机制。

SEA：Shared Ethernet Adapter 基于物理以太网适配器和虚拟以太网适配器创建，用来桥接外部网络和 PowerVM 内部虚拟交换机，这样可以实现虚拟客户端分区 VIOS 中的物理以太网适配器的虚拟使用，实现与外部网络的通信。

vSCSI：Virtual SCSI 让 VIO 服务器拥有适配器和磁盘，VIO 服务器可以分割磁盘并把部分磁盘（或整个磁盘）通过 vSCSI 协议提供给 LPAR，让客户机 LPAR 认为自己拥有完整的引导磁盘。

2.2 Virtual I/O Server 介绍

PowerVM 虚拟化功能部件的关键组件之一是 VIOS(Virtual IO Server)虚拟 I/O 服务器，是一个特殊定制化的 AIX 操作系统的逻辑分区，可以提供存储 IO 和网络资源的虚拟化，从而使得各个逻辑分区通过 VIOS 来共享这些物理资源。下面了解规划一个 VIOS 需要考虑哪些因素。

2.2.1 Virtual I/O Server 的最小资源需求

下表介绍了安装一个 VIOS 的最小资源需求：

资源	需求
HMC 或者 IVM	HMC 或者 IVM 作为 PowerVM 的管理器，是创建分区，管理分区时候必须的
存储适配器	VIOS 分区必须有一个磁盘适配器用以连接存储(外置或者内置)来安装操作系统
磁盘	安装 VIOS 的磁盘至少 30G
以太网适配器	创建 SEA 必须，可以是多个组成聚合，实

	现冗余或者提高带宽
内存	至少 1G
处理器	至少 0.1 颗处理器
PowerVM 版本	Express、Standard、Enterprise 至少一种

注：基于 CPU 和内存的分配，建议如下，CPU 分配 1 颗，内存分配 4G

2.2.2 VIOS 配置的考虑要素

我们在配置 VIOS 时候，还需要考虑如下一些要素：

- VIOS 是一个专用的分区，仅仅为客户端分区提供虚拟 IO 操作，其他应用程序不能运行在 VIOS 上
- VIOS 尽量保证足够多的 CPU 和内存资源，以免给客户端分区带来 IO 延迟的影响
- 通过 vSCSI 从 VIOS 映射给客户端分区使用的 LV 或者 ISO 文件只有单路径
- 目前只有以太网适配器可以作为 SEA 使用
- VIOS 不支持 IP 转发
- VIOS 中的虚拟适配器最大值可以是从 2 到 65536 中的任意值，但是如果你把虚拟适配器的最大值设置大于 1024，那么分区可能会启动失败或者需要更多的内存来管理这些虚拟适配器。所以通常我们设置虚拟设配器的最大值不要超过 1024。
- VIOS 支持客户端分区包括 AIX、IBM i、Linux，不同的产品支持不同的版本，在使用之间，建议查看产品手册以确认。

2.3 规划 VIOS 环境

PowerVM 为 VIOS 的配置提供了很大的灵活性，很多的选择，所以有多种不同的方法来规划实施虚拟化环境。根据具体的需求，选择一个最合适的规划很重要。

2.3.1 VIOS 的硬件规划

在规划设计 PowerVM 之前，需要了解所有的需求以确保硬件资源，软件，PowerVM 的 license 满足需求，考虑系统的可访问性，可用性，扩展性，性能，安全，可恢复性。

比如你需要 NPIV 技术，那么 Power 服务器上就需要配置 8Gb 的光纤卡，同时所连接的交换机也要支持 NPIV 特性；如果你需要 LPM 活动分区迁移，那么就需要 PowerVM 的企业版 license。

2.3.2 VIOS 的资源规划

VIOS 的 CPU 和内存的多少取决于很多因素，比如是否是高速率的物理卡(8Gb 光纤卡，10Gb 网卡)，多少个客户端分区，是否采用 NPIV 技术，客户端分区上的应用对 IO 的需求等等，需要在真实环境下不断的观察，然后选择一个最合适的值。对于初始值，建议如下：

中低端的服务器比如 CPU 少于 8 core，运行的分区较少，且无 8Gb 光纤卡和 10Gb 网卡，CPU 初始可以设置为 0.5，内存设置为 2G。

配置较高的服务器，如果不使用 NPIV 且无 10Gb 网卡，CPU 初始设置为 1，内存设置为 4G

配置较高的服务器，如果使用 NPIV 或者 10Gb 网卡，CPU 初始设置为 1，内存设置为 6G 以上。

2.3.3 VIOS 冗余性的选择

一个 Power 服务器中可以配置 1 个 VIOS，2 个 VIOS 或者多个 VIOS，配置多少取决于实际的需求，1 个 VIOS 对于客户端分区来说没用冗余，如果 VIOS 故障，所有的客户端分区不可用；2 个或者多个 VIOS 可以为客户端分区提供冗余，在一个 VIOS 故障或者维护的情况下，保证客户端分区仍然可用。

2.3.4 VIOS 网络和存储的考虑

VIOS 主要为客户端分区提供存储资源和网络资源的虚拟化，所以在规划 VIOS 的时候需要格外考虑网络和存储的资源，通常来讲，在一个 VIOS 里面至少规划 2 块网卡做 etherchannel 实现冗余，2 块光纤卡通过多路径软件做冗余。

除了硬件层面的考虑，我们还需要考虑虚拟化层面，比如虚拟网络有多少个 VLAN，VLAN 是否做 trunk；虚拟存储是采用 vSCSI，还是采用 NPIV，下表列出了网络和存储需要考虑的所有要素：

网络组件	存储组件
物理以太网卡 Physical Ethernet	物理光纤卡
虚拟以太网卡 Virtual Ethernet	虚拟光纤卡
共享以太网卡 SEA	Virtual SCSI
VLANs	SAN 存储
Etherchannel or Link Aggregation	MPIO 多路径软件

2.3.5 VIOS 中 Slot 号以及设备命名的考虑

在一个 PowerVM 环境中，一个好的命名规范对于后期的维护或者故障排除来说非常重要，可以节省大量的时间。

对于虚拟适配器的槽位号 Slot Number，建议如下：

虚拟网卡，VIOS 和 VIOC 从 10 开始，

vSCSI，从 30 开始，且 VIOS 和 VIOC 保持一致，对于一个双 VIOS 环境来说，第一个 VIOS 使用偶数，第二个 VIOS 使用奇数，比如第一个 VIOS 使用 30，第二个 VIOS 使用 31，那么对应的 VIOC 的两个 vSCSI 适配器使用 30 和 31，依次往下类推

VIOS1 **vhost**	**VIOS2** **vhost**	**VIOC** **vscsi**
30	31	30/31
32	33	32/33
34	35	34/35

NPIV，从 70 开始，且 VIOS 和 VIOC 保持一致，对于一个双 VIOS 环境来说，第一个 VIOS 使用偶数，第二个 VIOS 使用奇数，比如第一个 VIOS 使用 70，第二个 VIOS 使用 71，那么对应的 VIOC 的两个 vSCSI 适配器使用 70 和 71，依次往下类推

VIOS1 **NPIV**	**VIOS2** **NPIV**	**VIOC** **NPIV**
70	71	70/71

72	73	72/73
74	75	74/75

对于通过 vSCSI 映射给客户端分区的磁盘，BACKING DEVICE 的命令建议如下：

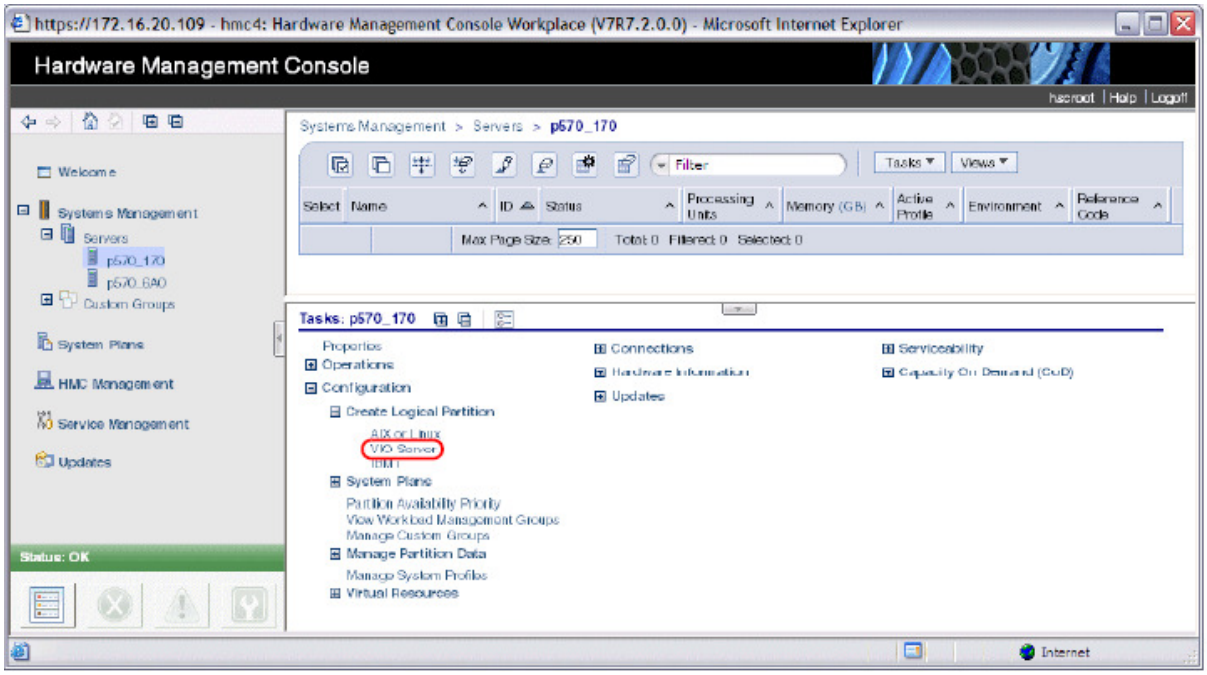
分区名_用途_数字，比如 hdisk1 给分区 app1 做 datavg 使用，那么这个 backing device 的名称可以是 app1_datavg_01。

3 VIOS 的安装升级

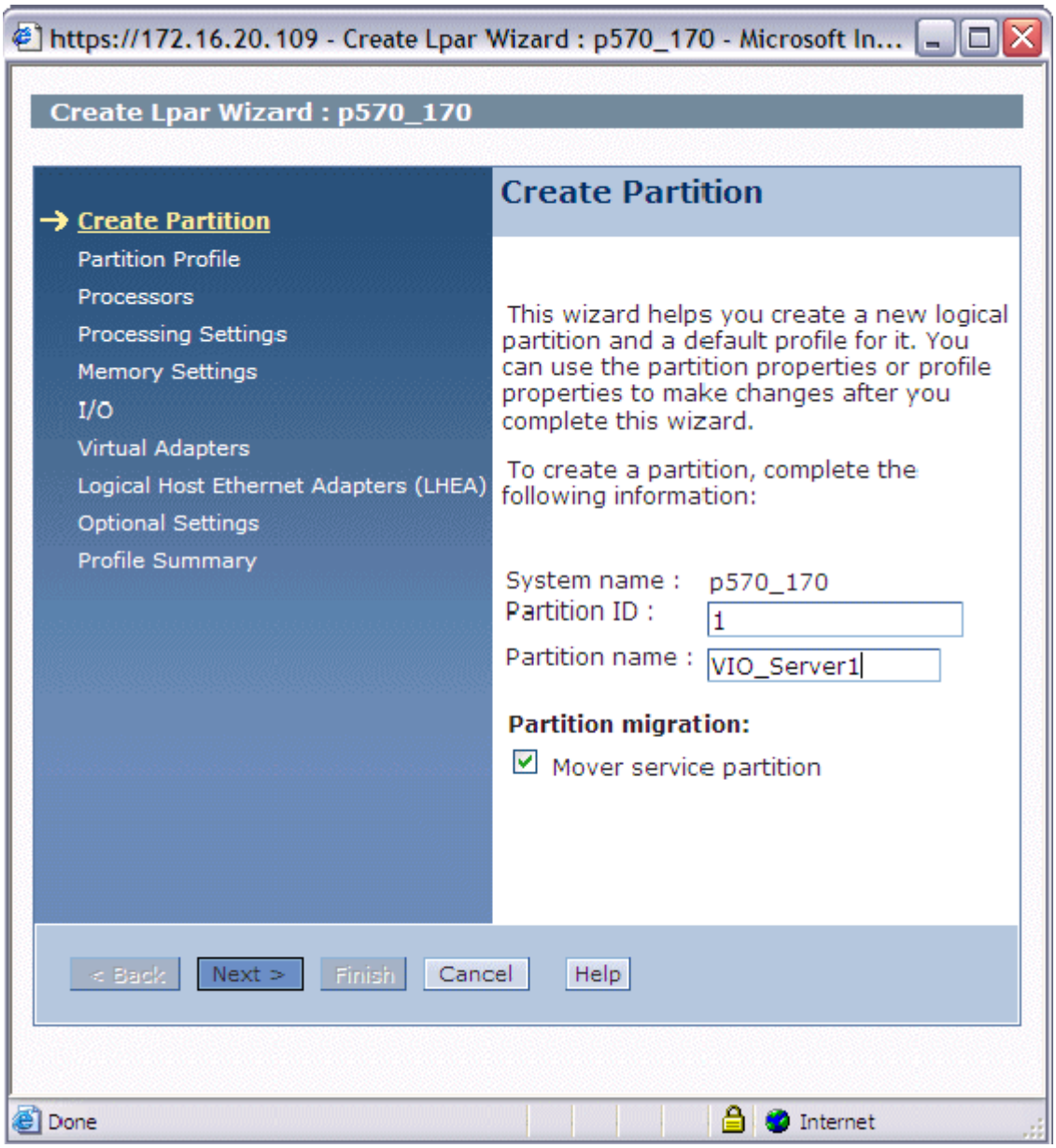
下面的内容主要描述如何创建、安装、配置 VIOS 分区

3.1 创建 VIOS 分区

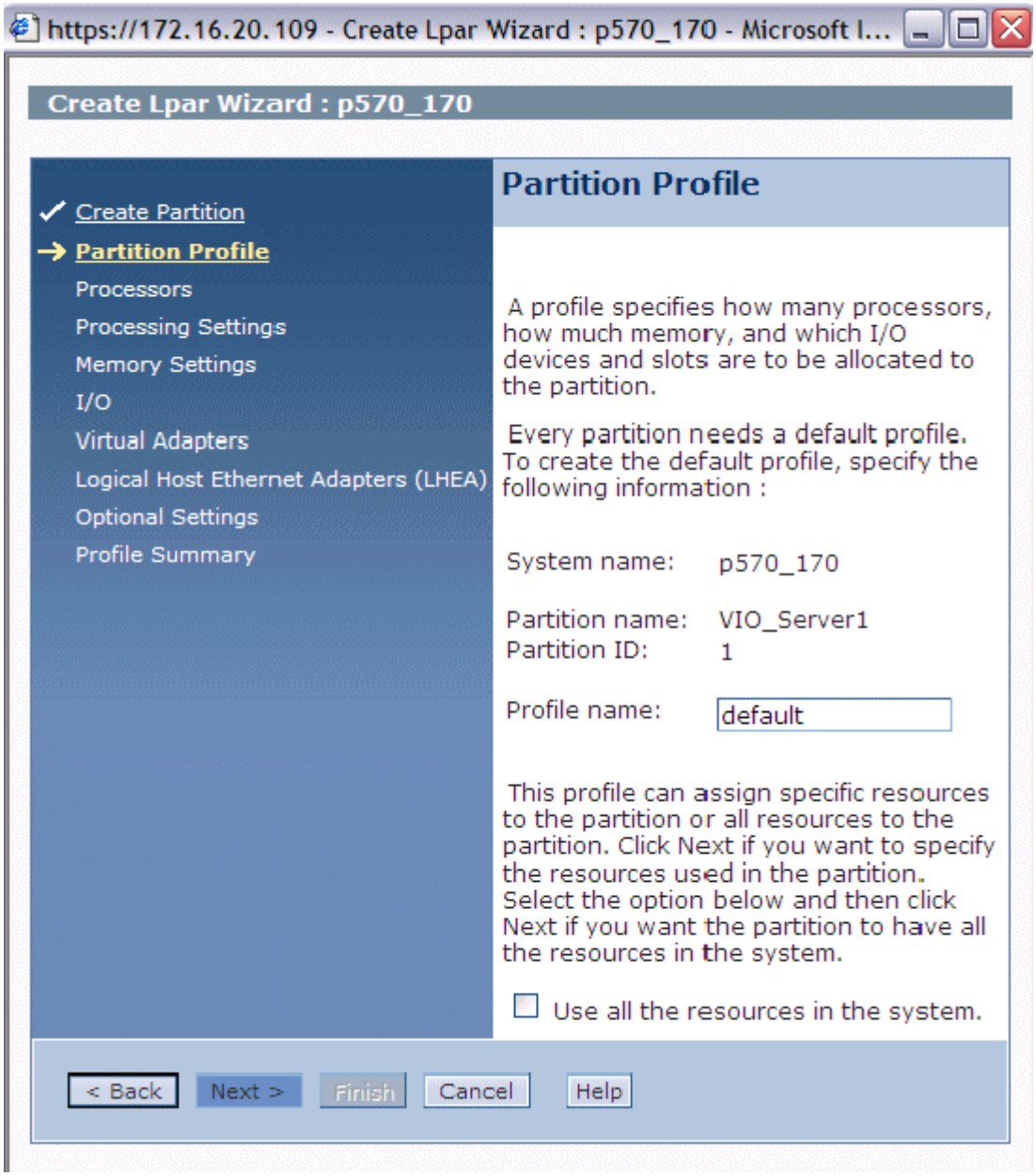
下面描述如何通过 HMC 创建一个 VIOS 分区，通过 HMC 上，然后选择“**Configuration -> Create Logical Partition -> VIO Server**”创建类型为 Virtual IO Server 类型的逻辑分区



输入分区 ID(如果分区 ID 没有特殊要求，使用 HMC 默认分配的即可)和分区名称，其中“Move service partition”记得要勾上来支持 VIOC 的 LPM，


https://172.16.20.109 - Create Lpar Wizard : p570_170 - Microsoft In...
Create Lpar Wizard : p570_170
Create Partition
Partition Profile
Processors
Processing Settings
Memory Settings
I/O
Virtual Adapters
Logical Host Ethernet Adapters (LHEA)
Optional Settings
Profile Summary
Create Partition
This wizard helps you create a new logical partition and a default profile for it. You can use the partition properties or profile properties to make changes after you complete this wizard.
To create a partition, complete the following information:
System name : p570_170
Partition ID : 1
Partition name : VIO_Server1
Partition migration:
Mover service partition
< Back
Next >
Finish
Cancel
Help
Done
Internet

输入概要文件的名称



3.1.1 CPU 设定

定义 CPU 资源，需要设置最小值，期望值和最大值。

Physical processing units:

PU(Processing Units) 是分区拥有的实际 CPU 处理能力，对于 PU(Processing Units)的值，最小建议设置成 0.1，期望值建议初始设置为 1，后续根据需要可以再做调整，最大值可以是期望值的两倍，比如 2。选择一个合适的最大值很重要，既要满足你在需要的时候通过 DLPAR 来调整 PU 的值来满足系统的需要，同时又不能设置的太大，因为设置的太大，PowerVM 的 Hypervisor 会消耗更多的内存资源。

Virtual processing units:

VP(Virtual processing)是分区系统里面可以看到的 CPU 个数，是 Hypervisor 可以调度的 CPU 个数。对于一个 uncapped 分区，VP 的值尤其重要，它可以允许在 CPU 资源足够的情况下，分区拥有 CPU 的处理能力超过 PU，而达到 VP 的值。设置 VP 的值需要慎重考虑，不能设置的太小，如果设置太小，达不到共享分区的效果，也不能设置的的太大，如果设置的太大，会造成太多的上下文切换(Context Switch)影响系统的性能。根据上面 PU 值的设定，建议 VP 的最小设置为 1，期望设置为 2，最大设置为 3 或者 4。

Capped or uncapped:

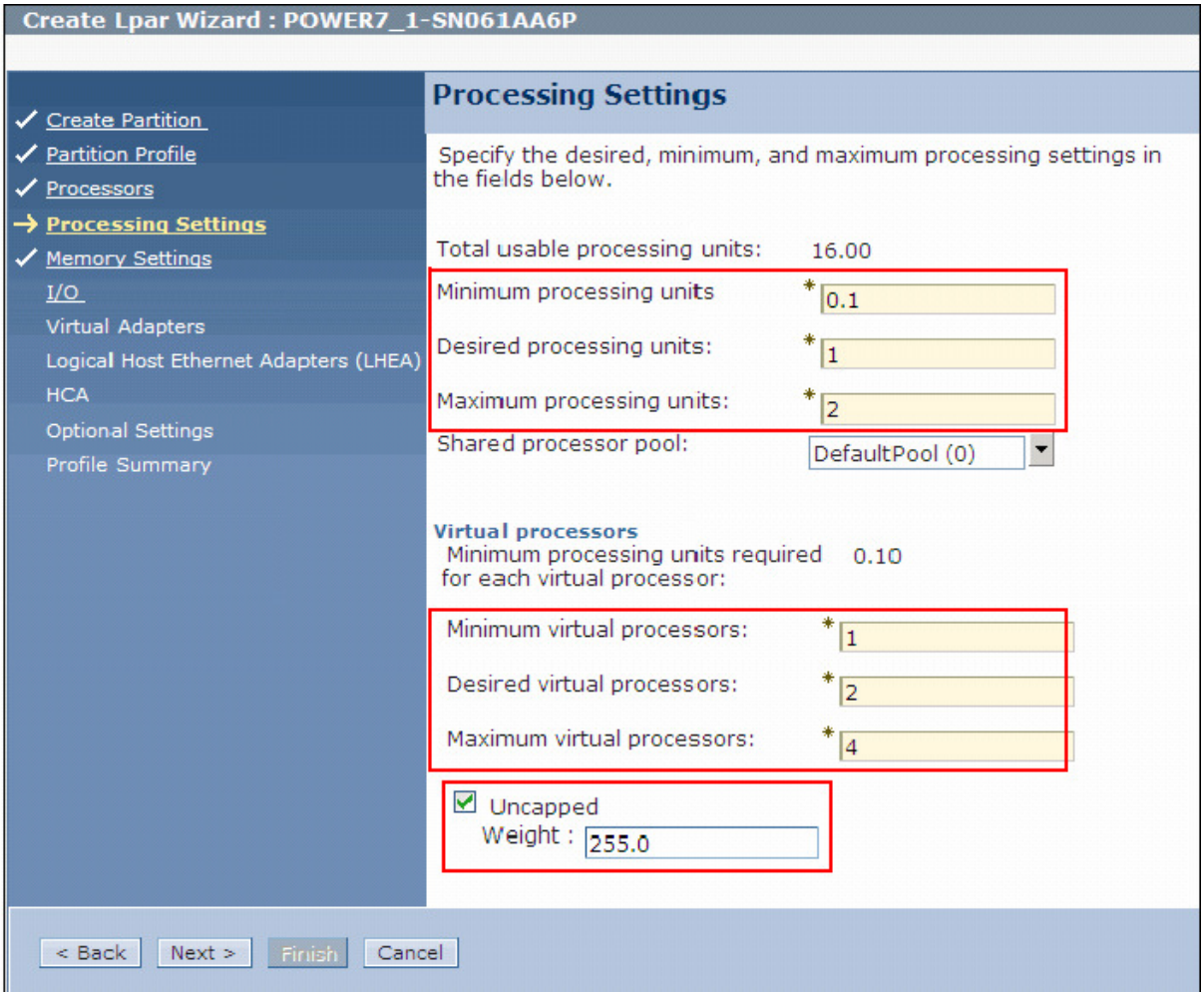
共享 CPU 的模式可以设置为 Capped (封顶)或者 Uncapped(不封顶)，capped 的意思是分区在运行过程中不能获取超过 PU 的期望值的处理能力；uncapped 的意思是分区在运行过程中可以获超过 PU 的处理能力，最大可以达到 VP 的值，如果 CPU 设置为共享模式，通常建议设置为 uncapped 模式。

Weight：

如果设置了 uncapped 选项，就需要设置一个 weight(权重)值，weight 值的大小从 0 到 255，其中 0 表示是 capped 分区，255 表示权重最大，在资源空闲的时候，优先级最高，通常我们设置 VIOS 的 weight 为最大，建议如下表：

Weight value	Usage
255	Virtual IO Server
200	Production
100	Pre-production
50	Development
25	Test

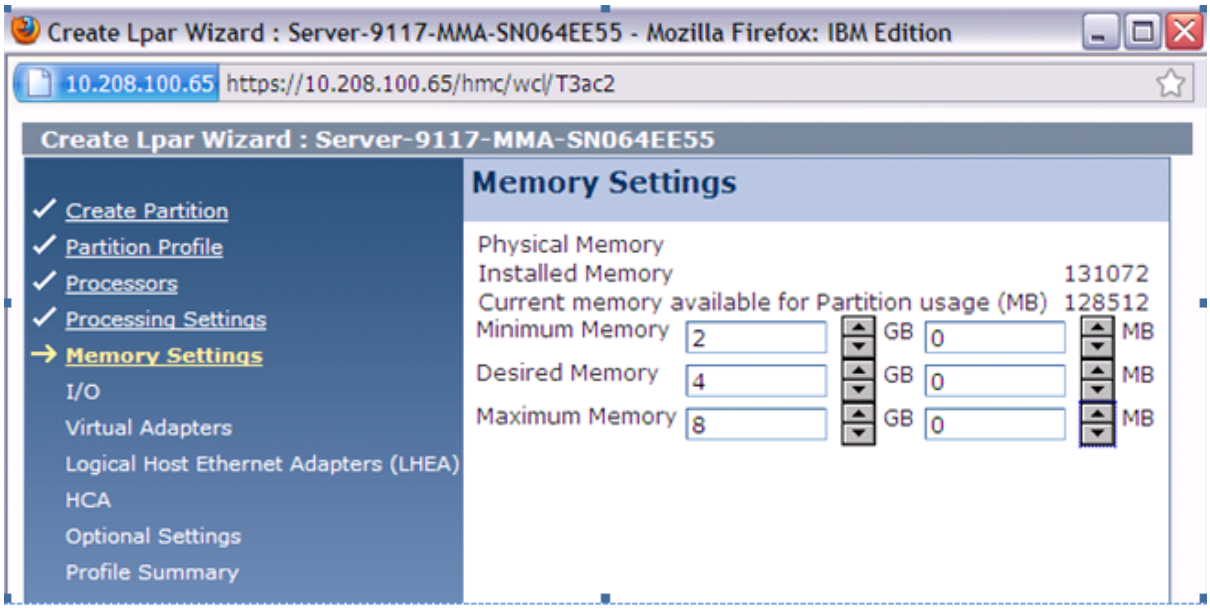
以下 VIOS 分区 CPU 设定的示例：



下面进入内存的设置

3.1.2 内存设定

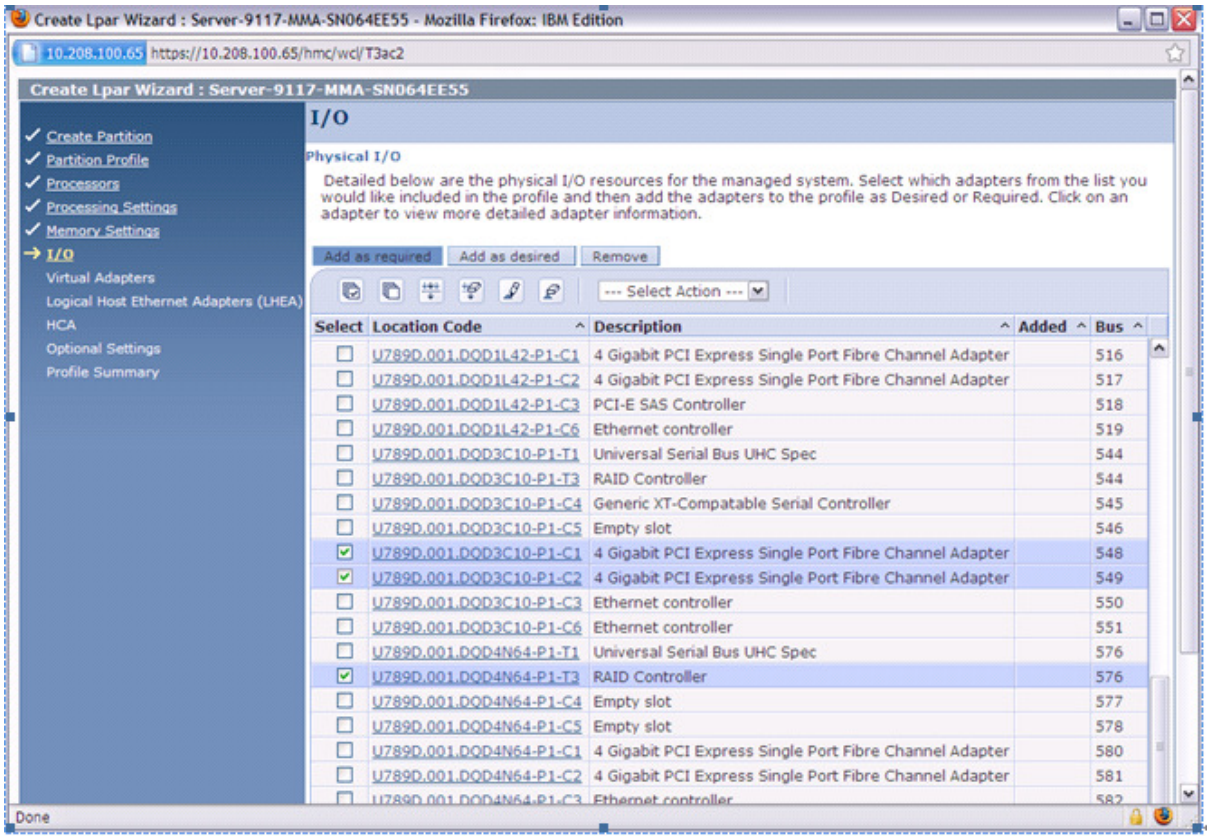
内存的设定与 CPU 一样，也需要三个值：最小值、期望值、最大值，也分为共享模式和独占模式，通过内存我们设置独占模式，初始值最小设置为 1G 或者 2G，期望设置为 4 或者 6G，最大设置为 8G



下面设置物理 IO 适配器

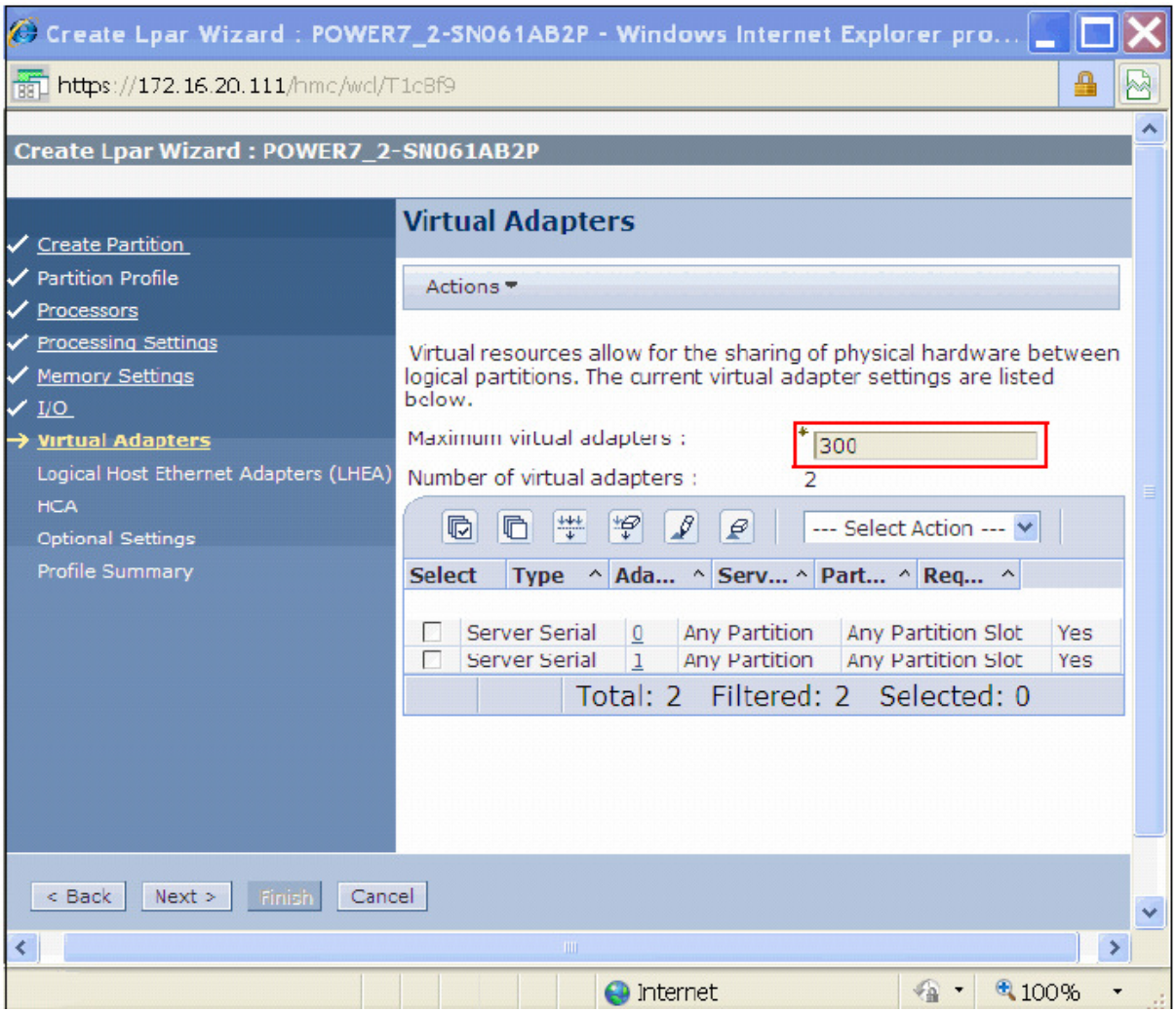
3.1.3 物理 IO 适配器设定

物理 IO 适配器的分配有两个值，必须(Required)和期望(Desired)，如果这个物理 IO 适配器给一个分区独占，且不需要给其他分区使用，设置为 required；如果这个物理 IO 适配器还需要通过 DLPAR 方式移动给其他分区使用(比如光驱所在的 IO 适配器)，那么设置为 Desired，根据规划分配对应的物理 IO 适配器。

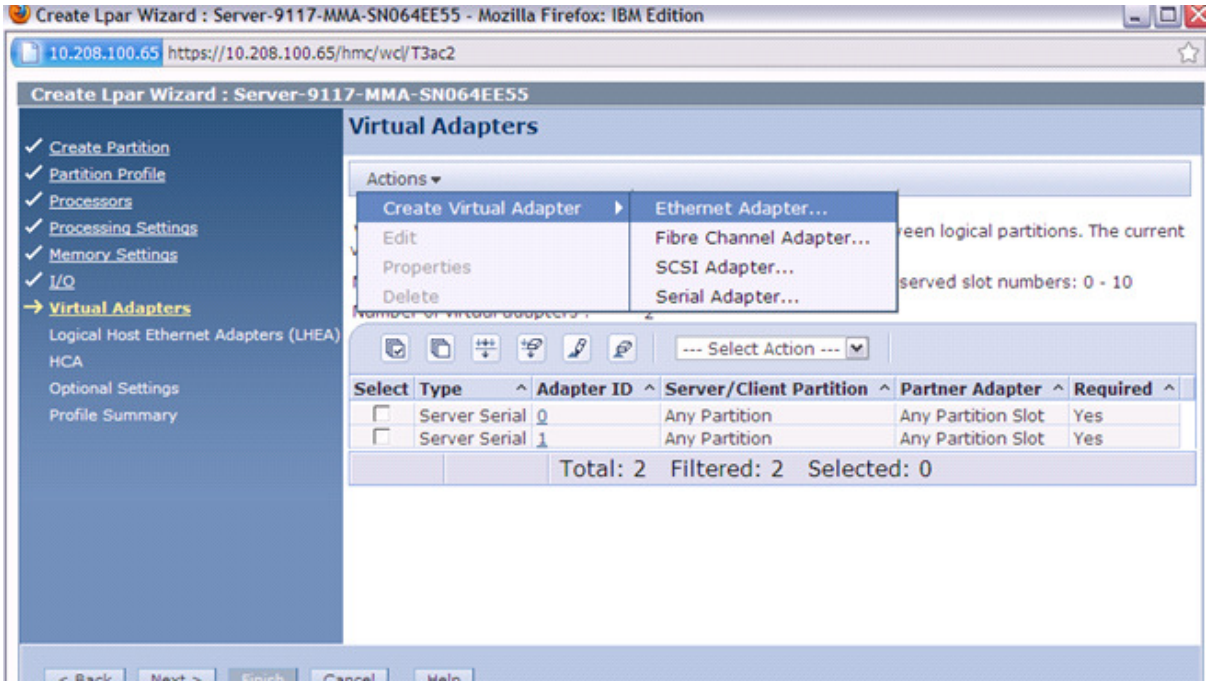


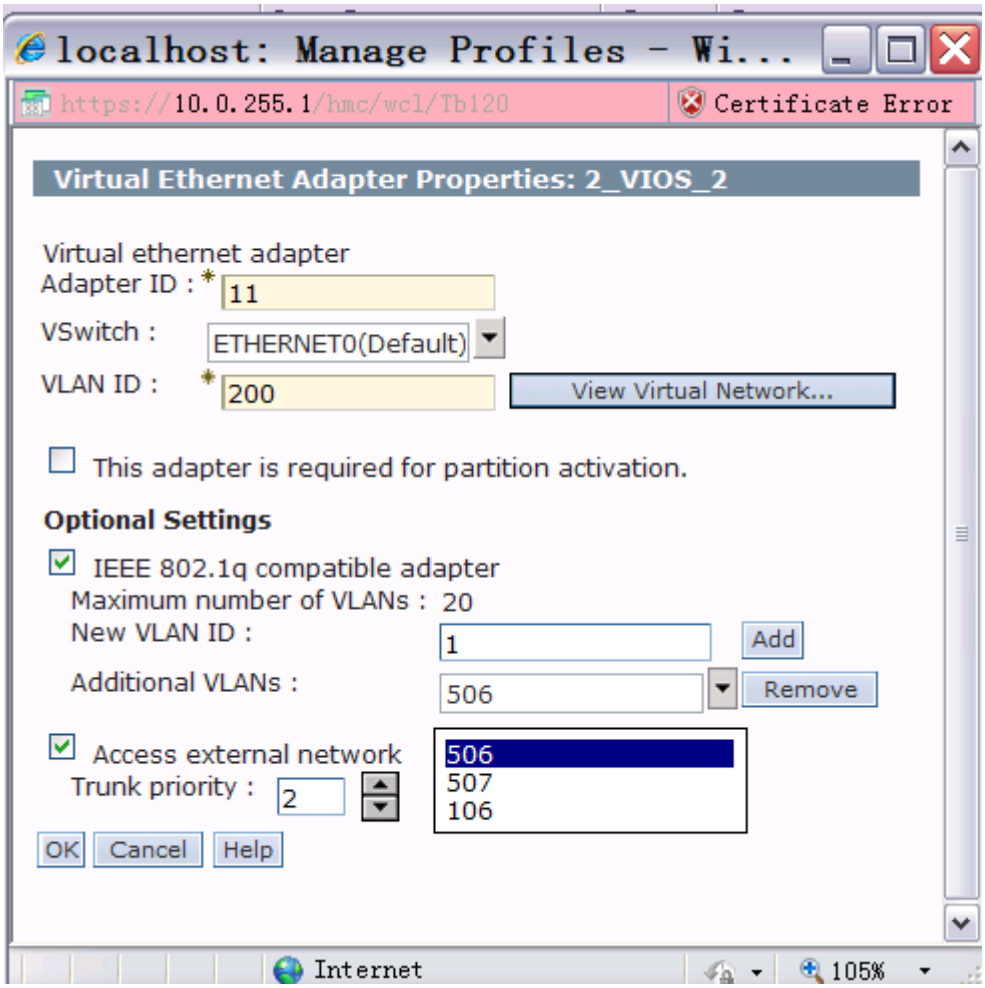
3.1.4 虚拟 IO 适配器设定

虚拟 IO 适配器的分配有两个值，必须(Required)和期望(Desired)，对于 VIOS 来讲，建议所有的虚拟 IO 适配器都设置为 Desired，同时修改虚拟适配器的最大个数，建议设置的大点，比如 1000



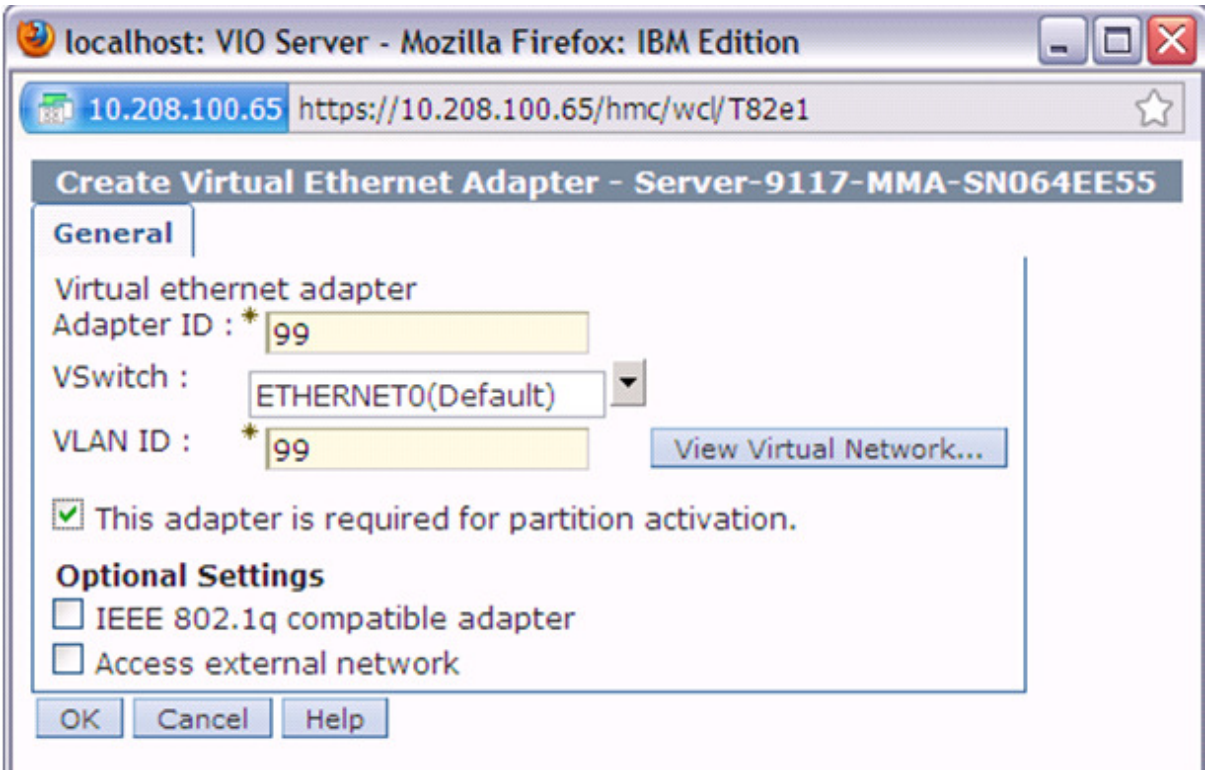
创建虚拟网卡：选择” Action→Create Virtual Adapter→Ethernet Adapter”：



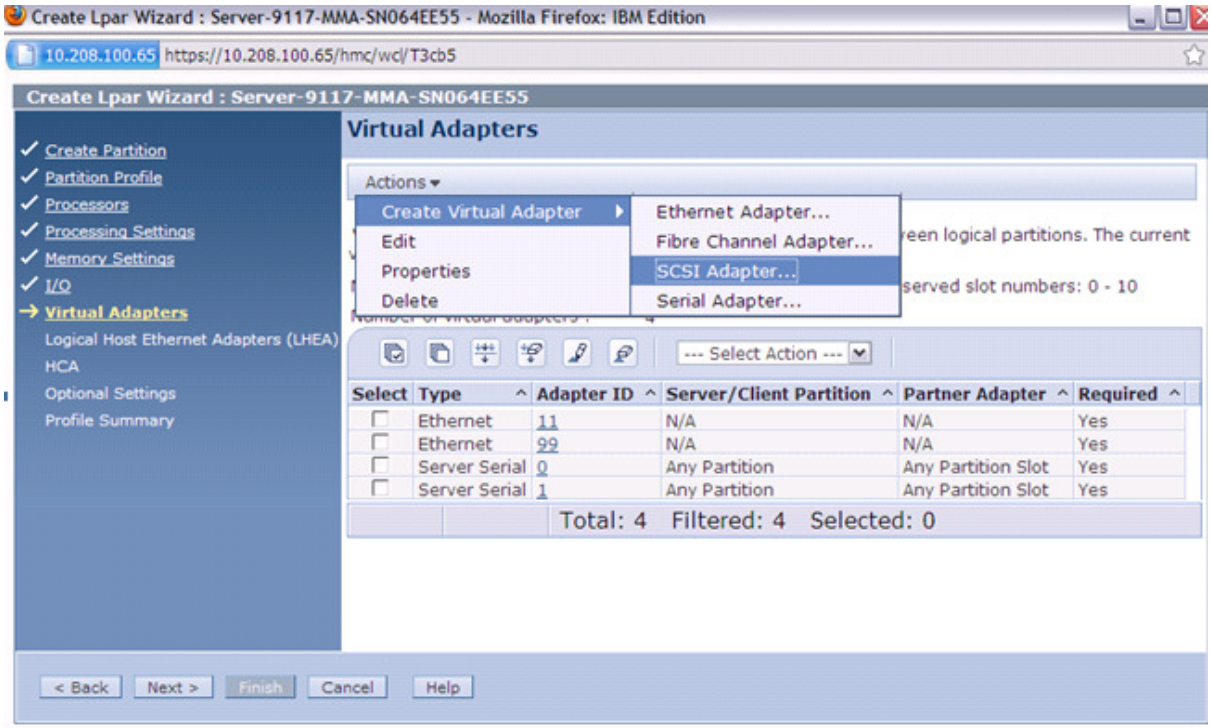


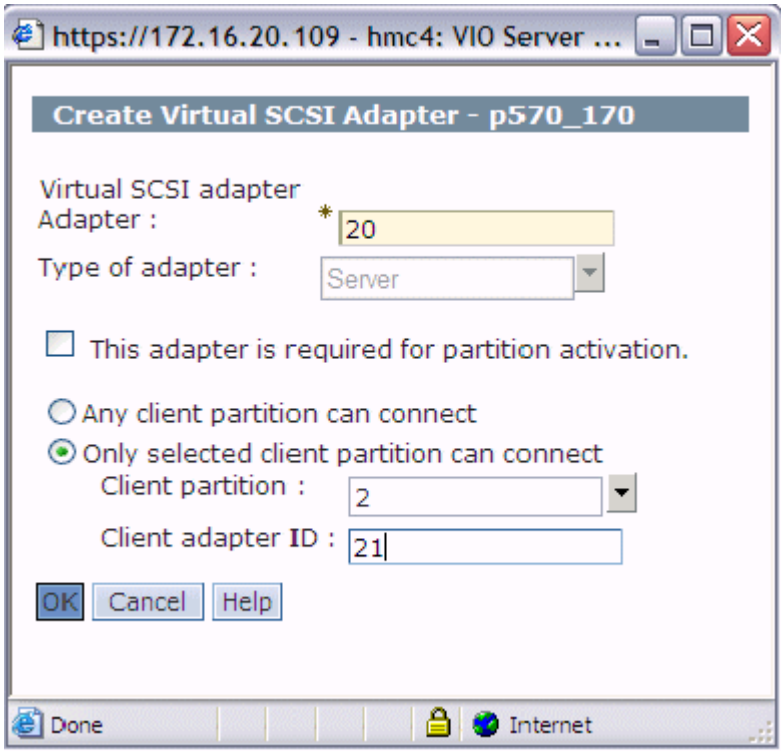
如上示例创建一个 Slot ID 为 11，VLAN ID 为 200，Additional VLAN 为 106,506,507，优先级为 2 的虚拟网卡，其中 Access external network(访问外部网络)需要勾上，表示可以做

SEA 使用；如果不要做 Trunk，则没有 Additional VLAN，即 IEEE 802.1q compatible adapter 不需要勾上；如果需要 Trunk，则需要配 Additional VLAN，且默认的 VLAN ID(PVID)要是一个闲置无用的 ID，且在外面连接的交换机上不存在这个 VLAN ID；优先级的设置跟 SEA 的冗余，双 VIOS 有关，优先级为 1 表示优先级最高，两个 VIOS 的优先级不能设置为一样，优先级高的表示分区对外的网络优先通过这个 VIOS。在双 VIOS 的配置中，每个 SEA 我们还需要创建一个额外的虚拟网卡作为控制通道(Control Channel)以控制 SEA 的切换，以免造成环路，通过这个网卡的 PVID 在两个 VIOS 上要设置成一样，没有 Trunk，不需要访问外部网络，比如以下，可以设置 VLAN ID 为 99 的虚拟网卡作为控制通道使用



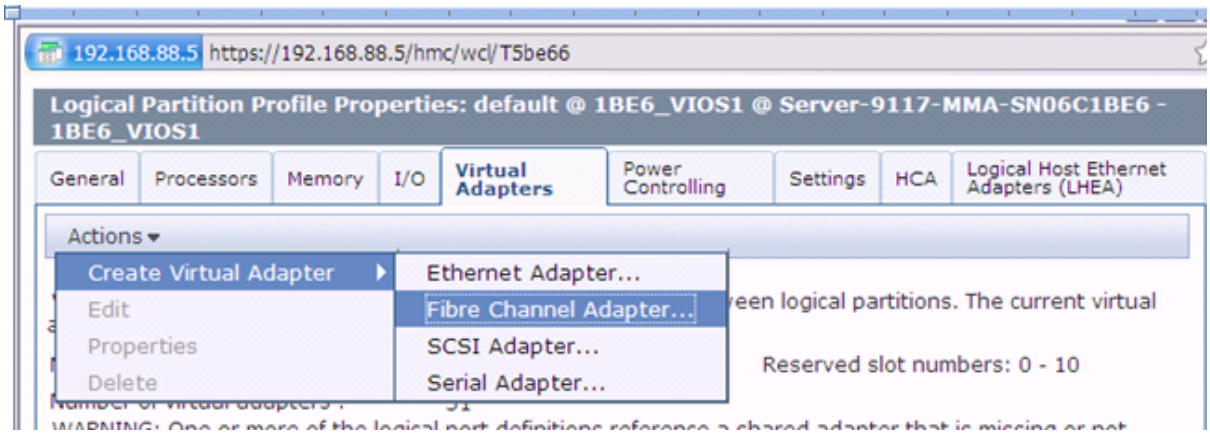
创建虚拟 SCSI：选择” Action→Create Virtual Adapter→Ethernet Adapter”：

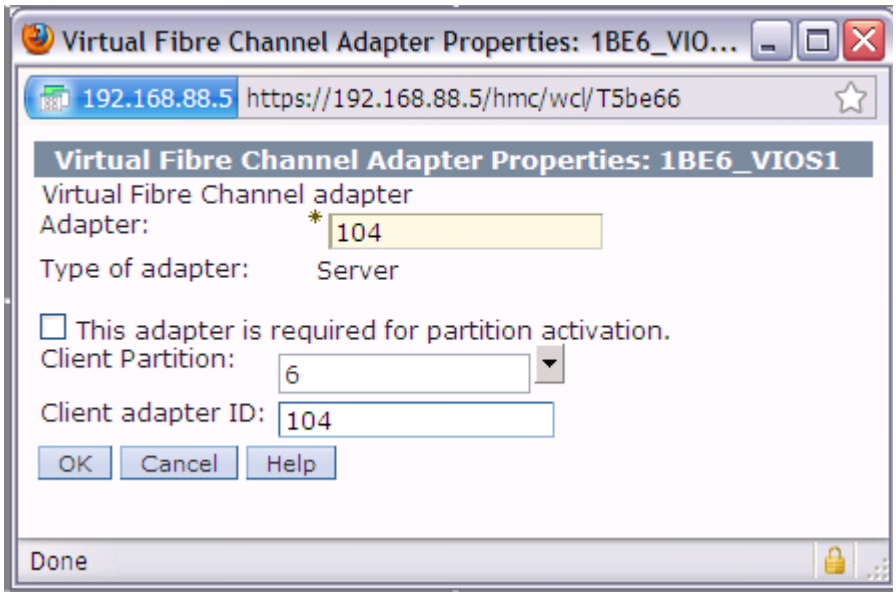




根据规划输入 vSCSI 的 Slot Number，对应的 Client 分区，由于此时客户端分区还没创建，所以在 Client partition 和 Client adapter ID 处需要手工输入相应的值，注意对于 VIOS 上虚拟适配器不要选择 Required。

创建虚拟光纤卡：选择” Action→Create Virtual Adapter→Fibre Channel Adapter”：





根据规划输入 VFC 的 Slot Number，对应的 Client 分区，由于此时客户端分区还没创建，所以在 Client Partition 和 Client adapter ID 处需要手工输入相应的值，注意对于 VIOS 上虚拟适配器不要选择 Required。这里表示 VIOS 上 Slot 为 104 的 VFC Server Adapter 对应分区号为 6 的分区的 VFC Client adapter。

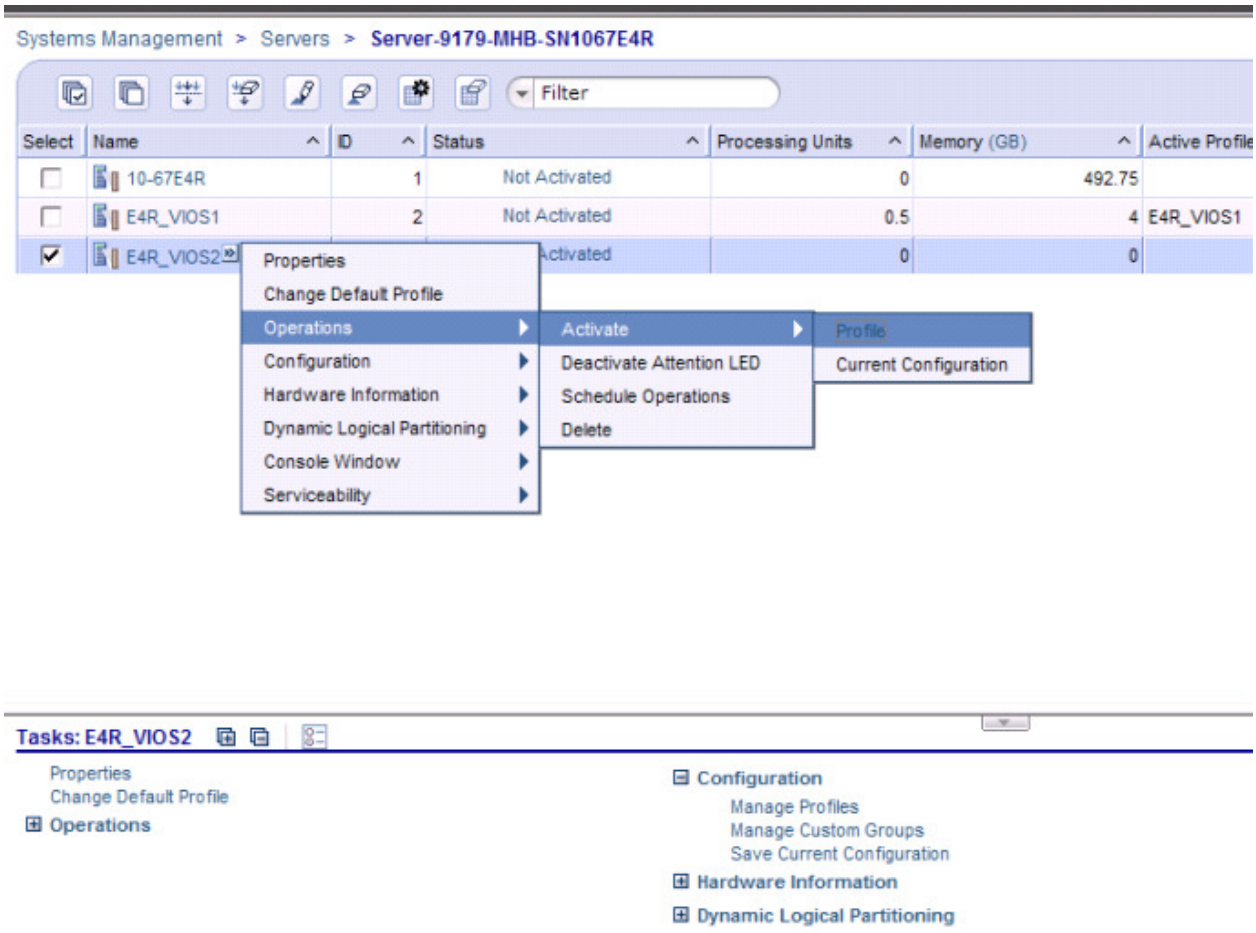
创建好以上的资源后，后面的选项选择默认即可，直到完成分区的创建。

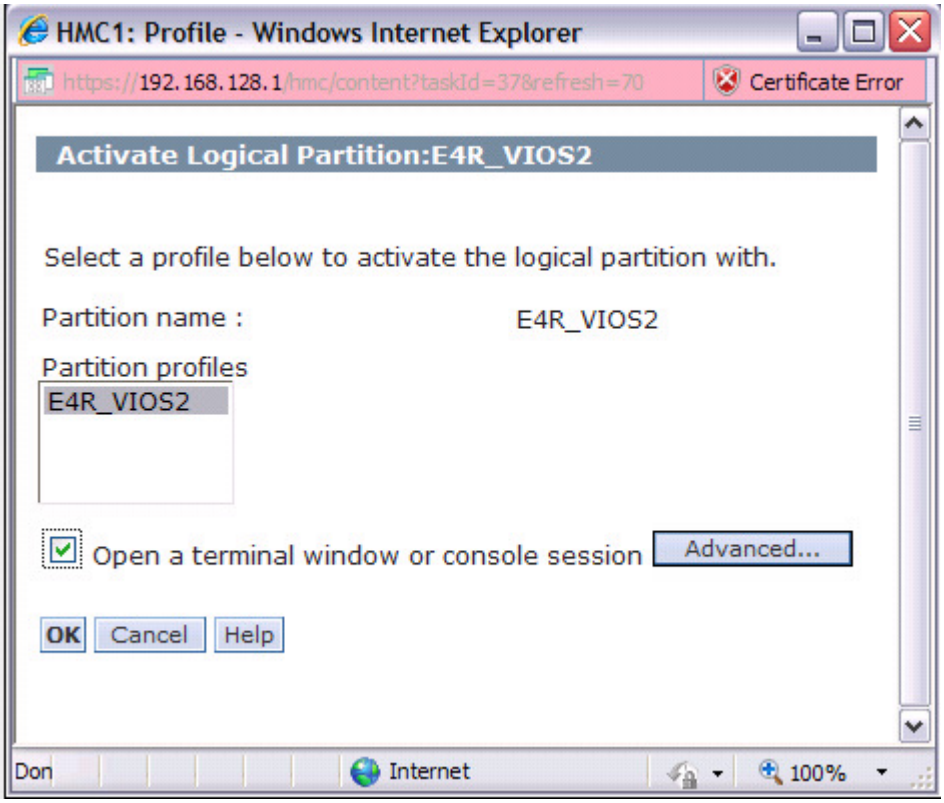
3.2 VIOS 安装

3.2.1 VIOS 的安装

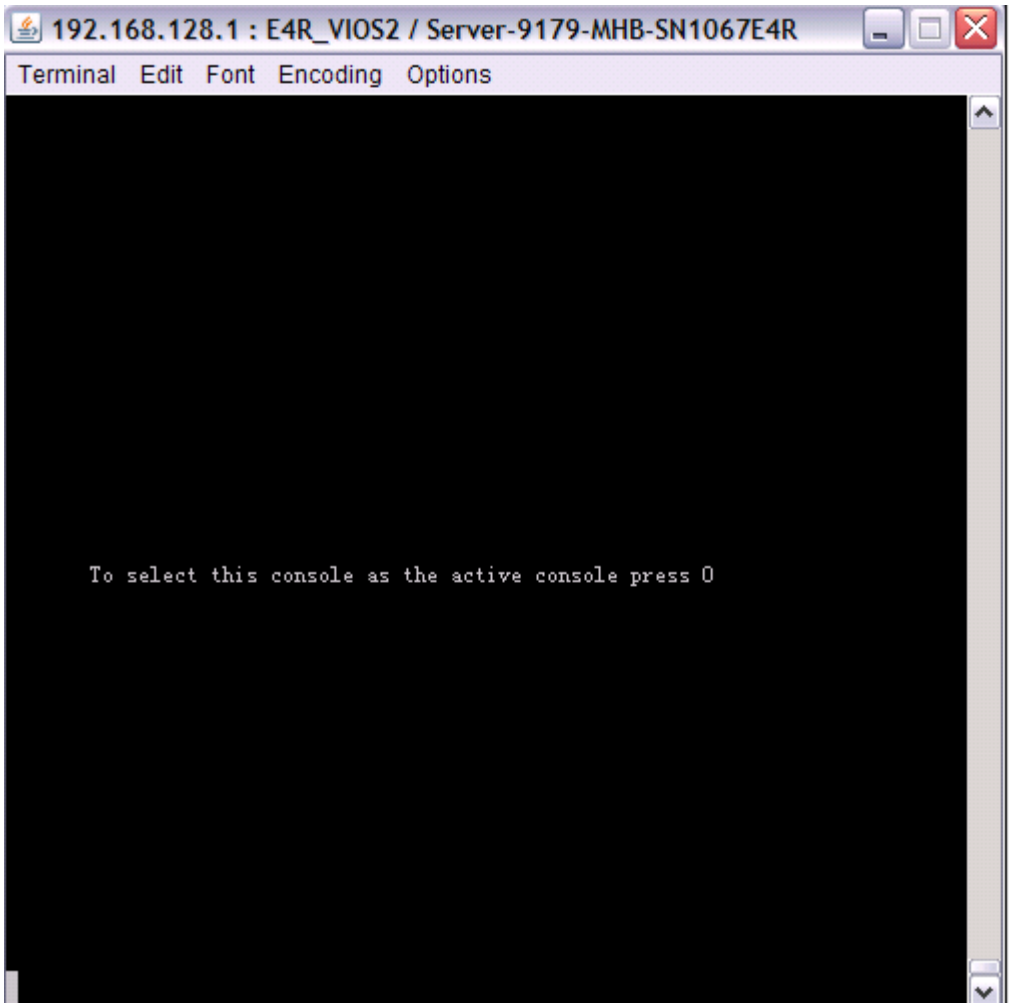
VIOS 一般安装在服务器的本地磁盘上（也可以安装在外置存储盘上，即 SANBOOT 方式），其安装方式与 AIX 系统安装相似，可以通过光盘来安装，也可以采用 NIM 或系统克隆方式来安装。在大批量 VIOS 部署时，建议采用 NIM 方式（需要具备 NIM 服务器和网络）来安装或系统克隆方式安装（通过 VIOS 自带的命令 alt_root_vg）。下面介绍通过光盘在服务器本地磁盘上安装 VIOS 并配置系统镜像的过程。

在 HMC 上激活 VIOS 分区，并 Open Terminal Windows 打开终端，

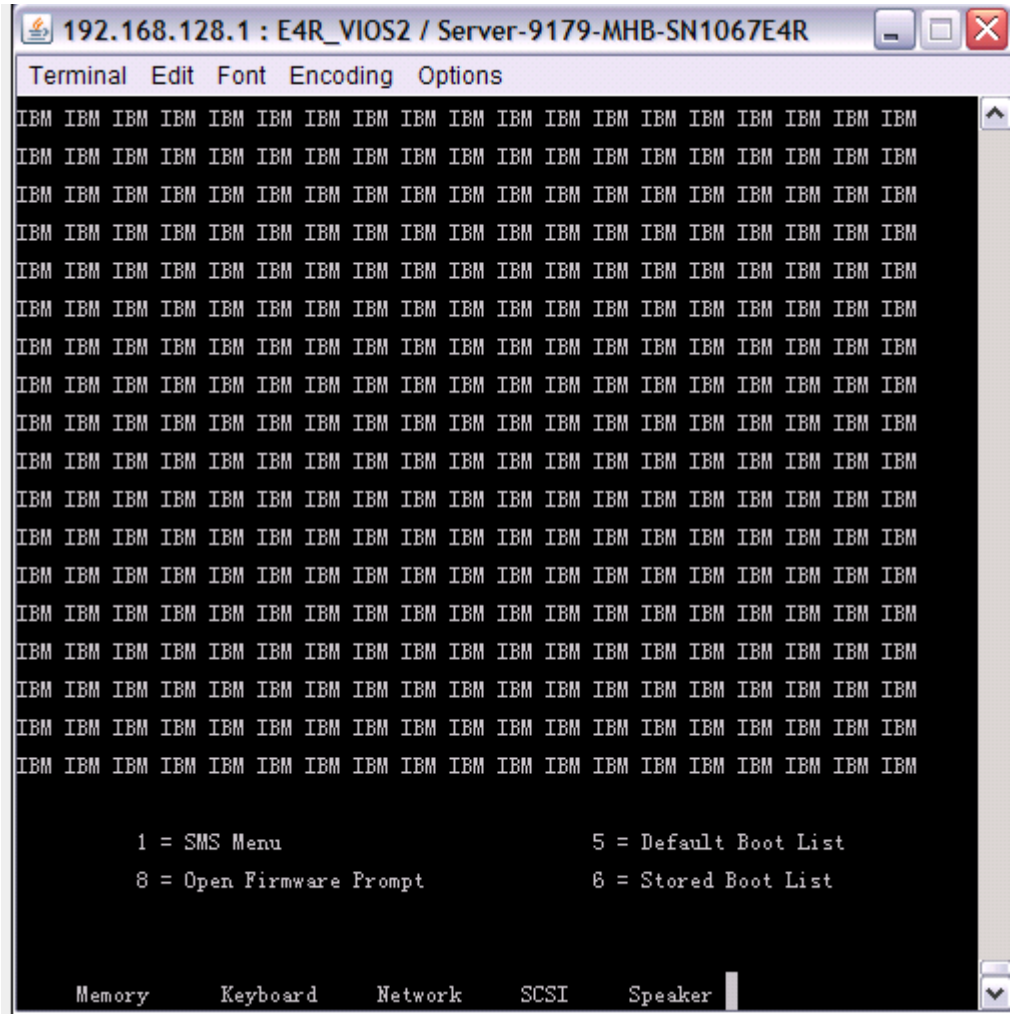



HMC1: Profile - Windows Internet Explorer
https://192.168.128.1/hmc/content?taskId=378refresh=70
Certificate Error
Activate Logical Partition:E4R_VIOS2
Select a profile below to activate the logical partition with.
Partition name :
E4R_VIOS2
Partition profiles
E4R_VIOS2
Open a terminal window or console session
Advanced...
OK
Cancel
Help
Don
Internet
100%

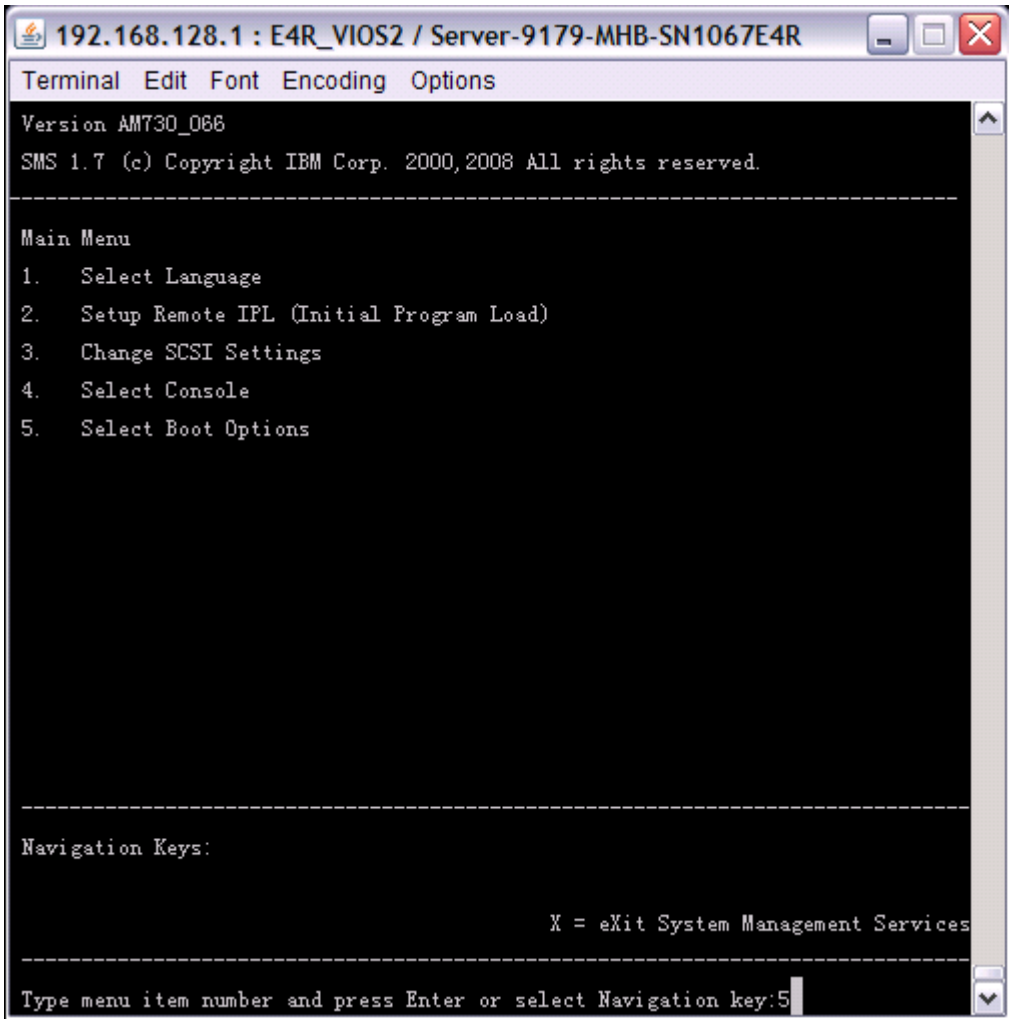
选择 OK


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
To select this console as the active console press 0

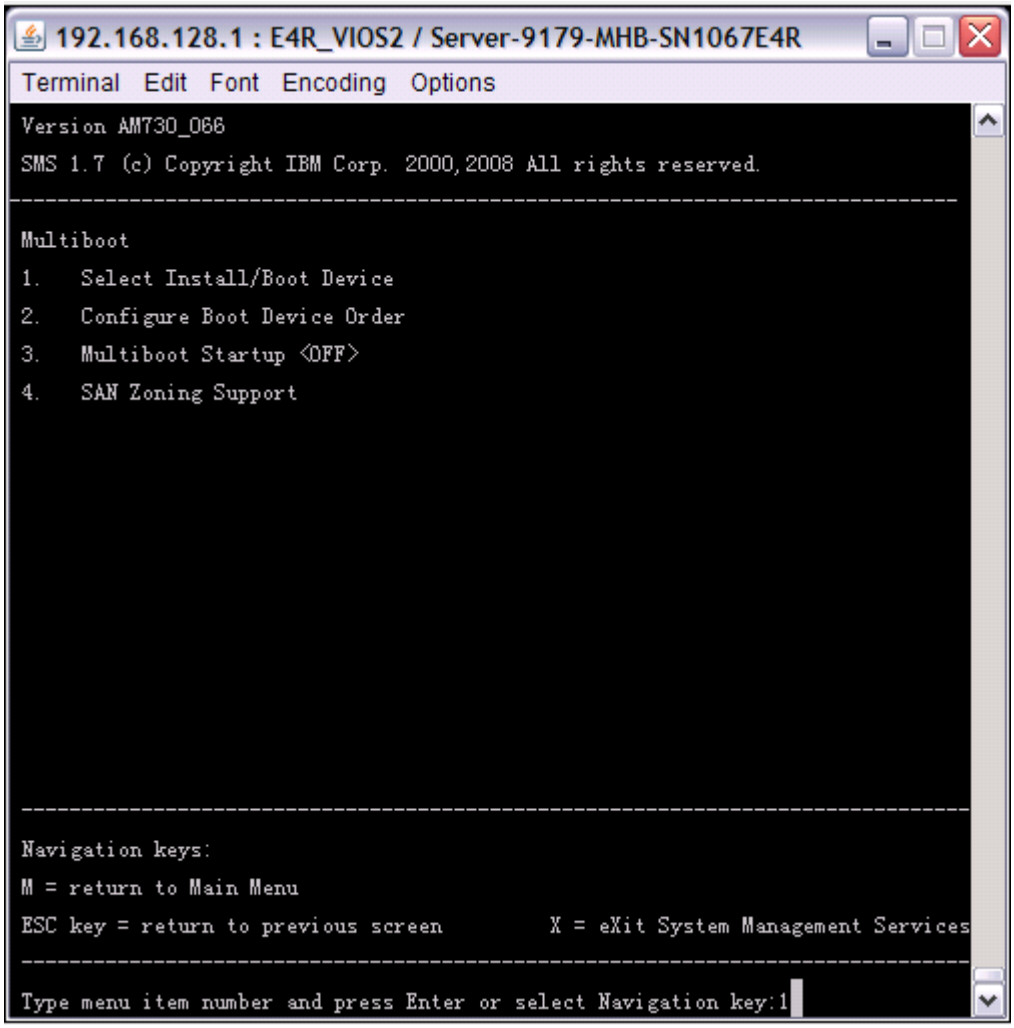
按 0 选择当前终端


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM IBM
1 = SMS Menu
5 = Default Boot List
8 = Open Firmware Prompt
6 = Stored Boot List
Memory Keyboard Network SCSI Speaker

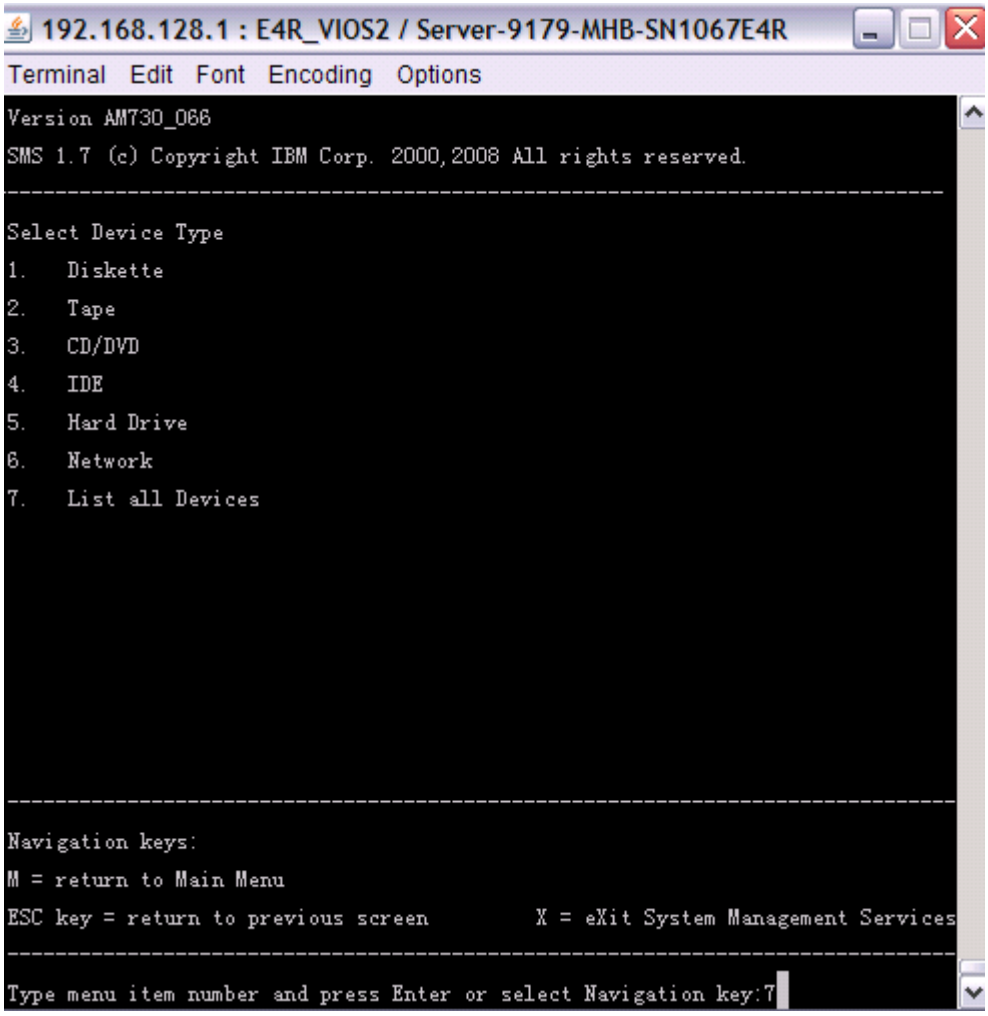
按 1 进入 SMS 菜单


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Version AM730_066
SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.
Main Menu
1. Select Language
2. Setup Remote IPL (Initial Program Load)
3. Change SCSI Settings
4. Select Console
5. Select Boot Options
Navigation Keys:
X = eXit System Management Services
Type menu item number and press Enter or select Navigation key:5

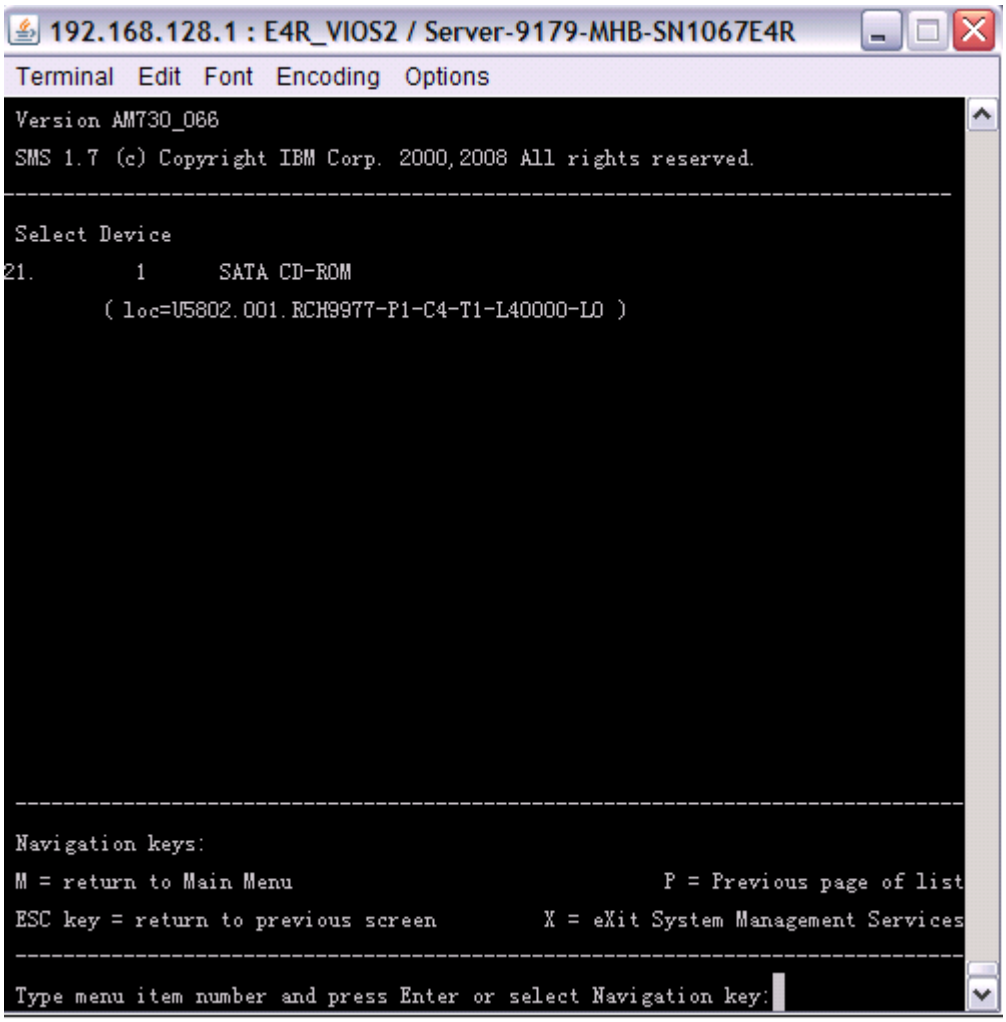
按 5 选择启动选项


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Version AM730_066
SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.
Multiboot
1. Select Install/Boot Device
2. Configure Boot Device Order
3. Multiboot Startup <OFF>
4. SAN Zoning Support
Navigation keys:
M = return to Main Menu
ESC key = return to previous screen
X = eXit System Management Services
Type menu item number and press Enter or select Navigation key:1

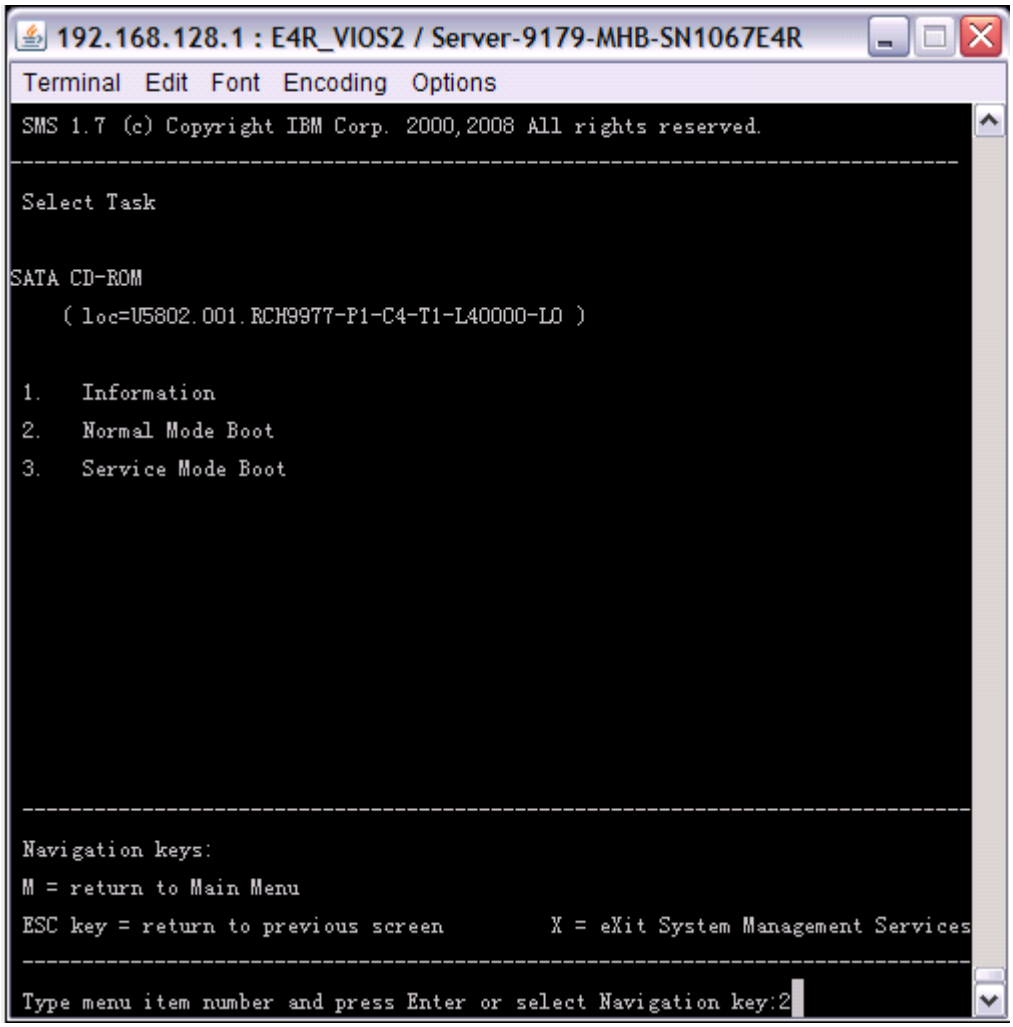
按 1 选择启动设备


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Version AM730_066
SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.
Select Device Type
1. Diskette
2. Tape
3. CD/DVD
4. IDE
5. Hard Drive
6. Network
7. List all Devices
Navigation keys:
M = return to Main Menu
ESC key = return to previous screen
X = eXit System Management Services
Type menu item number and press Enter or select Navigation key:7

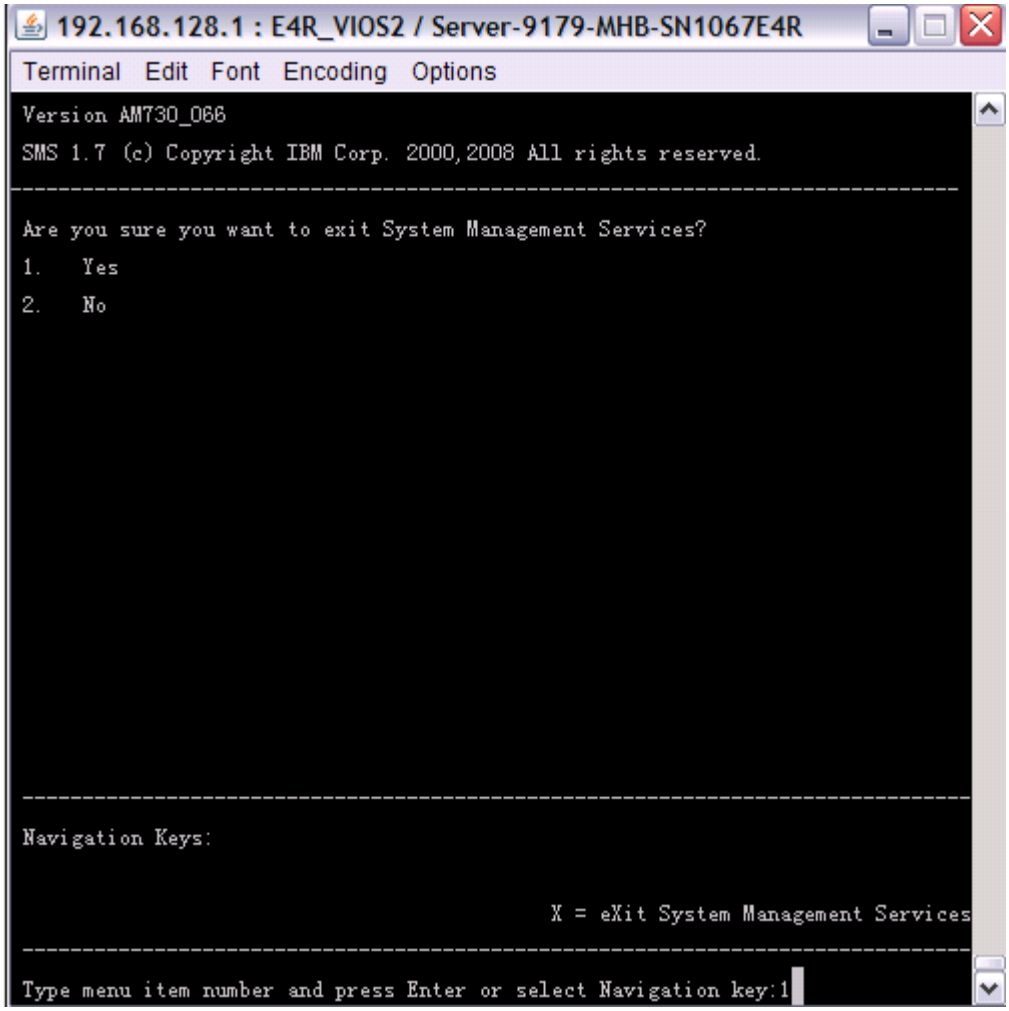
按 7 列出所有设备


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Version AM730_066
SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.
Select Device
21. 1 SATA CD-ROM
(loc=U5802.001.RCH9977-P1-C4-T1-L40000-L0)
Navigation keys:
M = return to Main Menu
P = Previous page of list
ESC key = return to previous screen
X = eXit System Management Services
Type menu item number and press Enter or select Navigation key:

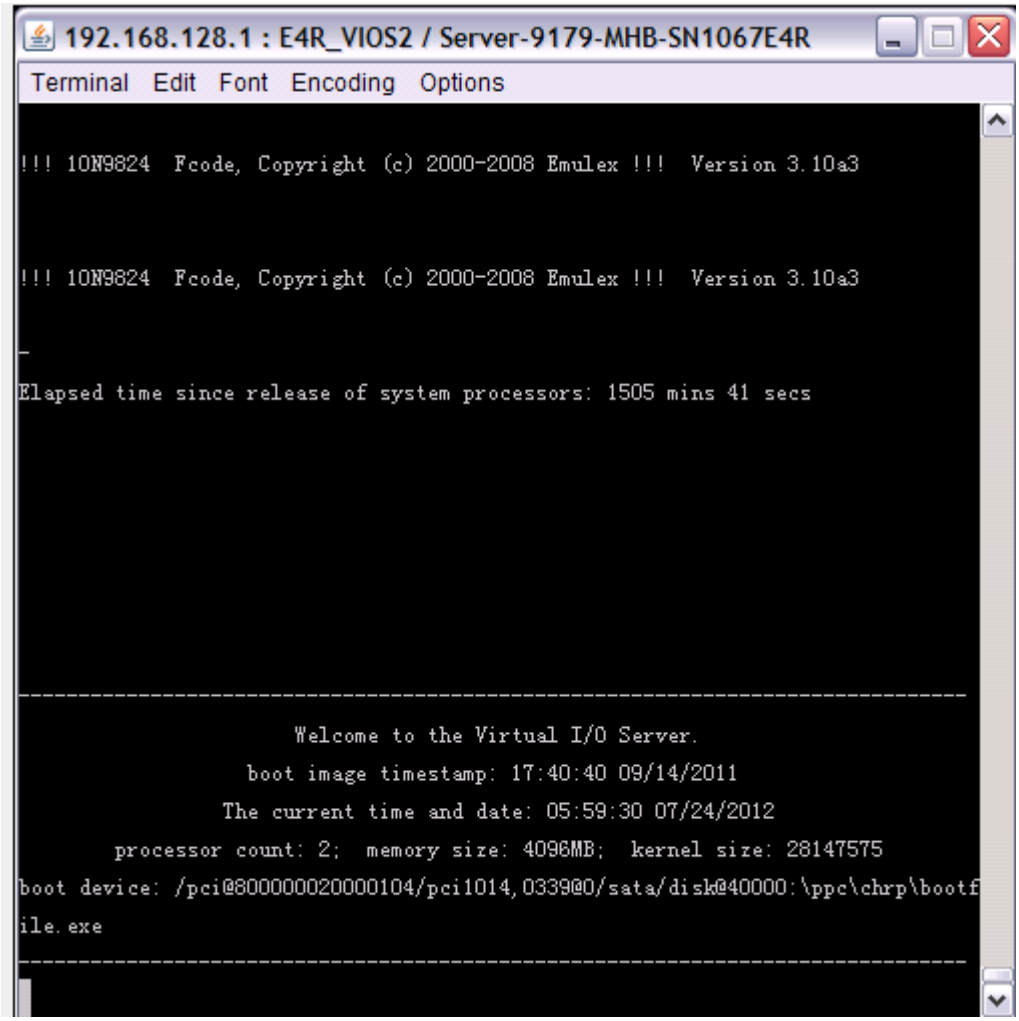
按 1 选择光驱



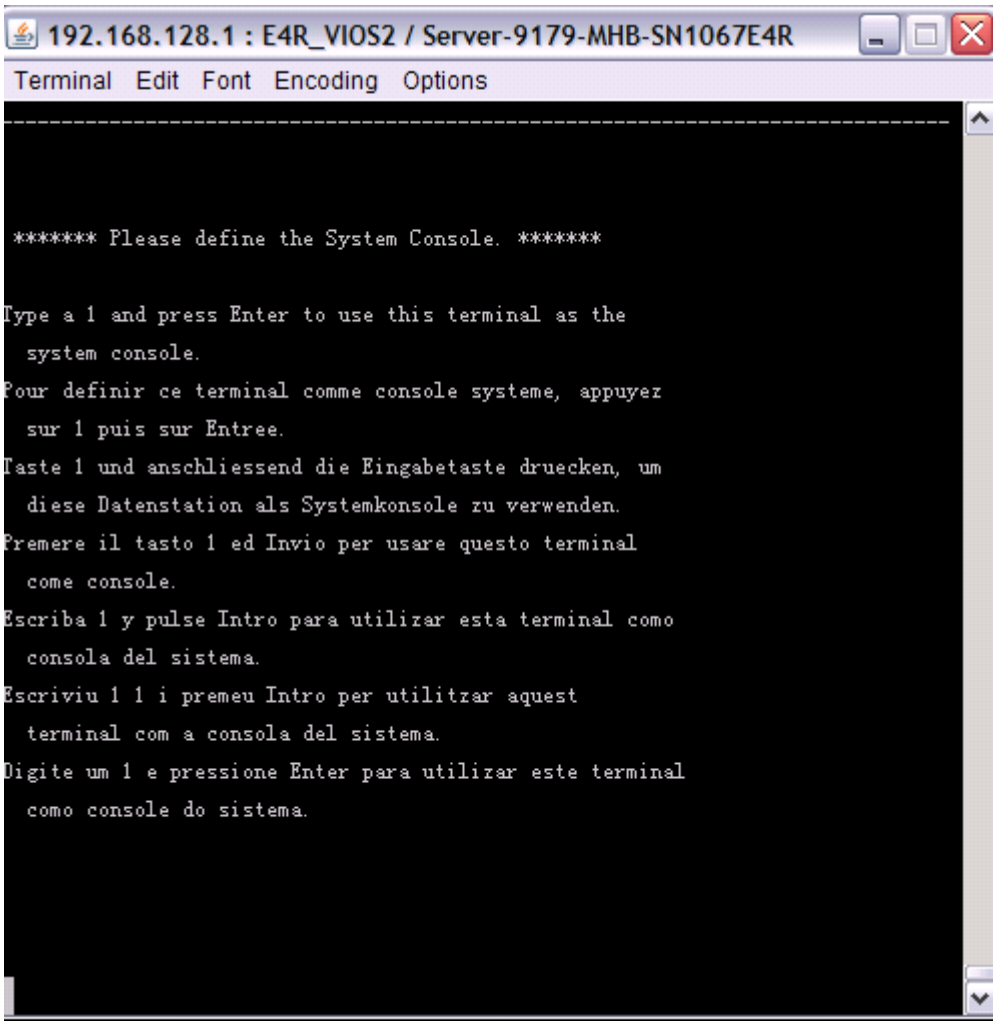
按 2 选择 Normal 模式



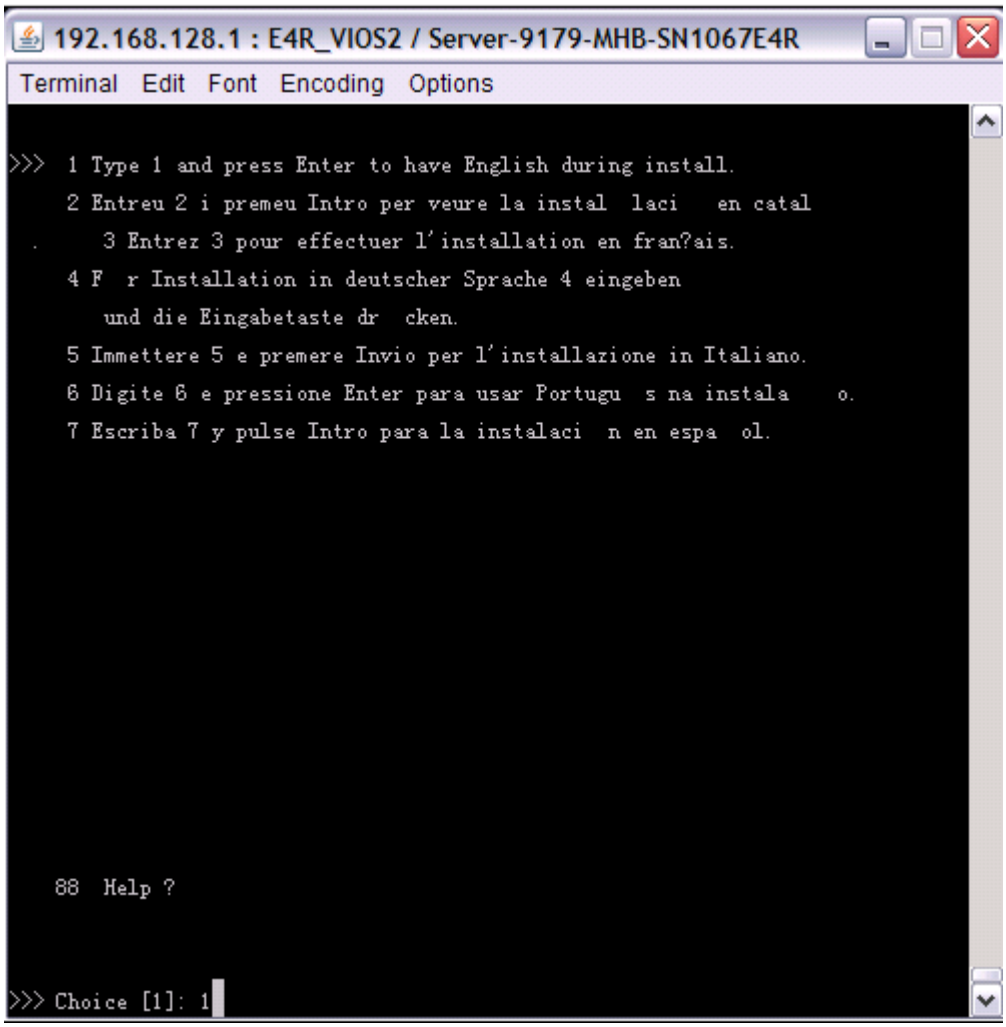
按 1 选择离开 SMS 菜单，开始安装


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
!!! 10N9824 Fcode, Copyright (c) 2000-2008 Emulex !!! Version 3.10a3
!!! 10N9824 Fcode, Copyright (c) 2000-2008 Emulex !!! Version 3.10a3
Elapsed time since release of system processors: 1505 mins 41 secs
Welcome to the Virtual I/O Server.
boot image timestamp: 17:40:40 09/14/2011
The current time and date: 05:59:30 07/24/2012
processor count: 2; memory size: 4096MB; kernel size: 28147575
boot device: /pci@800000020000104/pci1014,0339@0/sata/disk@40000:\ppc\chrp\bootf
ile.exe

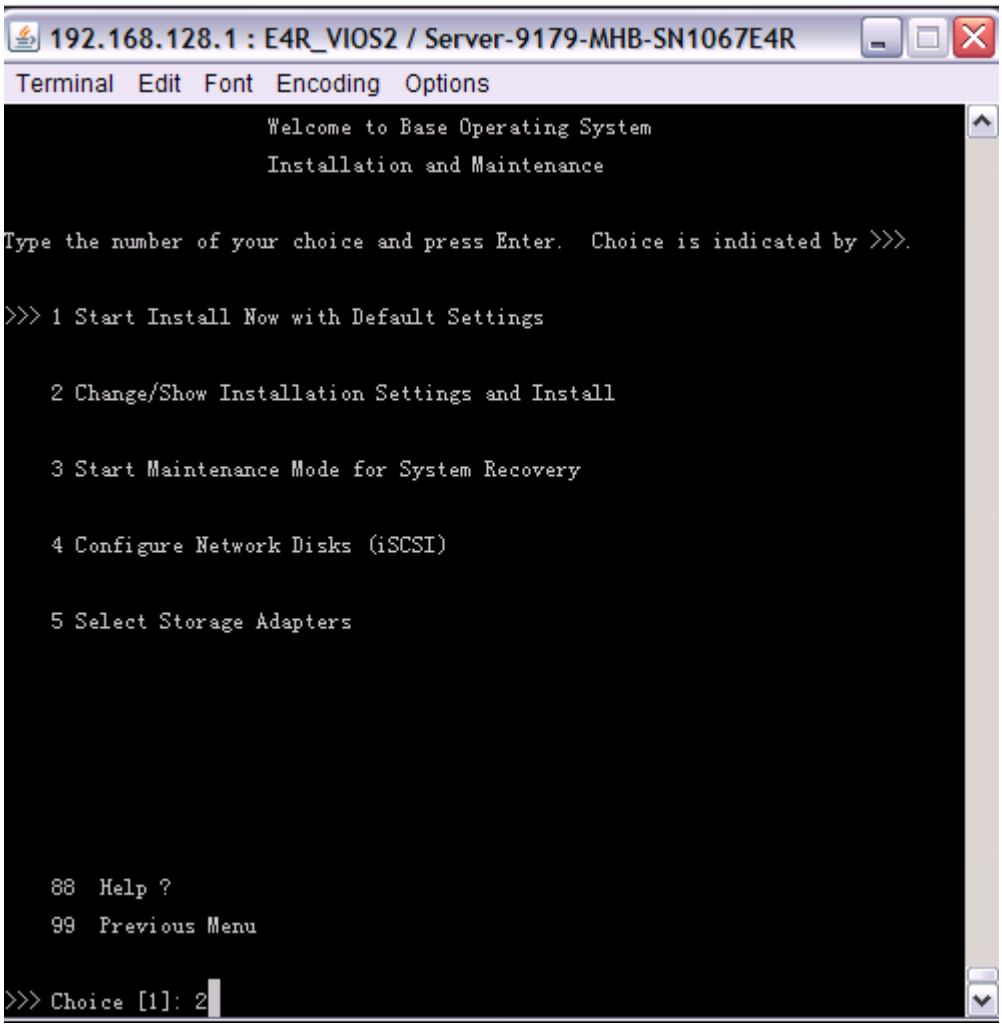
光盘开始引导


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
******* Please define the System Console. *******
Type a 1 and press Enter to use this terminal as the
system console.
Pour definir ce terminal comme console systeme, appuyez
sur 1 puis sur Entree.
Taste 1 und anschliessend die Eingabetaste druecken, um
diese Datenstation als Systemkonsole zu verwenden.
Premere il tasto 1 ed Invio per usare questo terminal
come console.
Escriba 1 y pulse Intro para utilizar esta terminal como
consola del sistema.
Escriviu 1 i premeu Intro per utilitzar aquest
terminal com a consola del sistema.
Digite um 1 e pressione Enter para utilizar este terminal
como console do sistema.

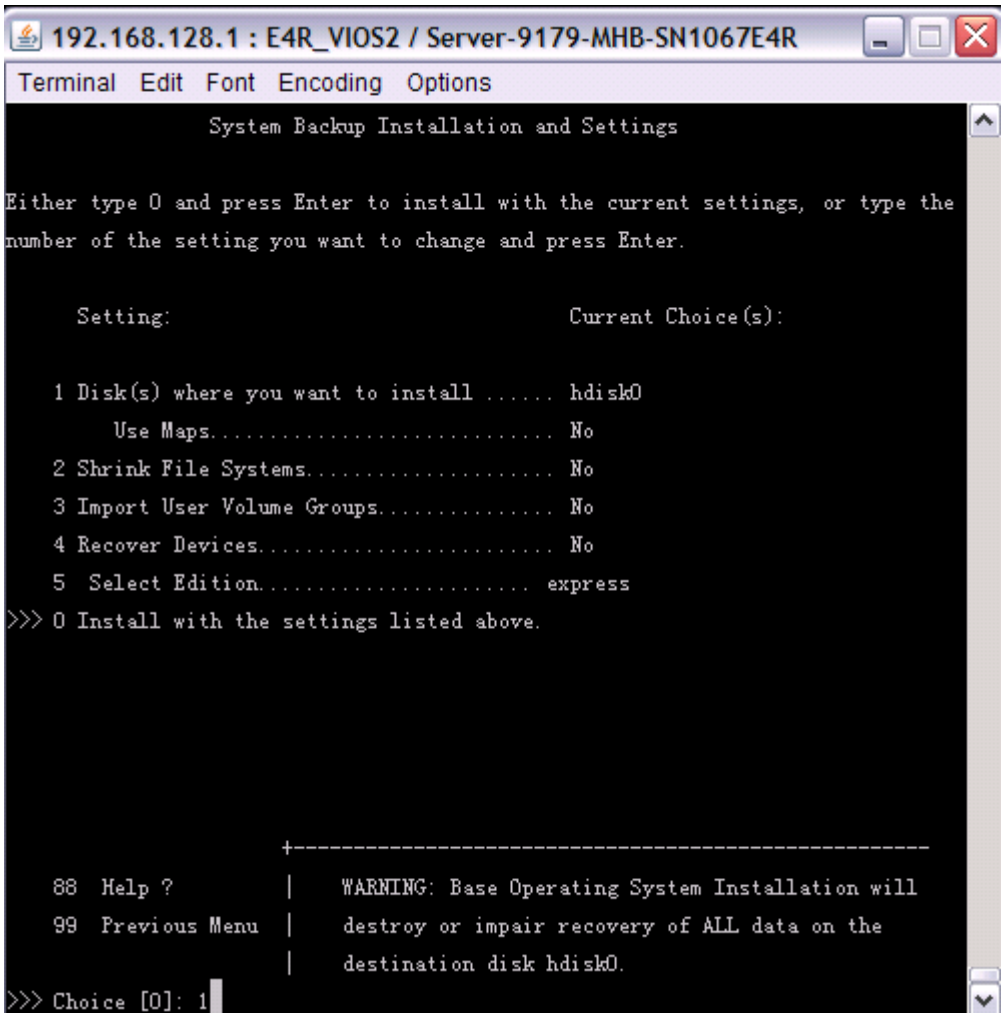
按 1 选择当前终端



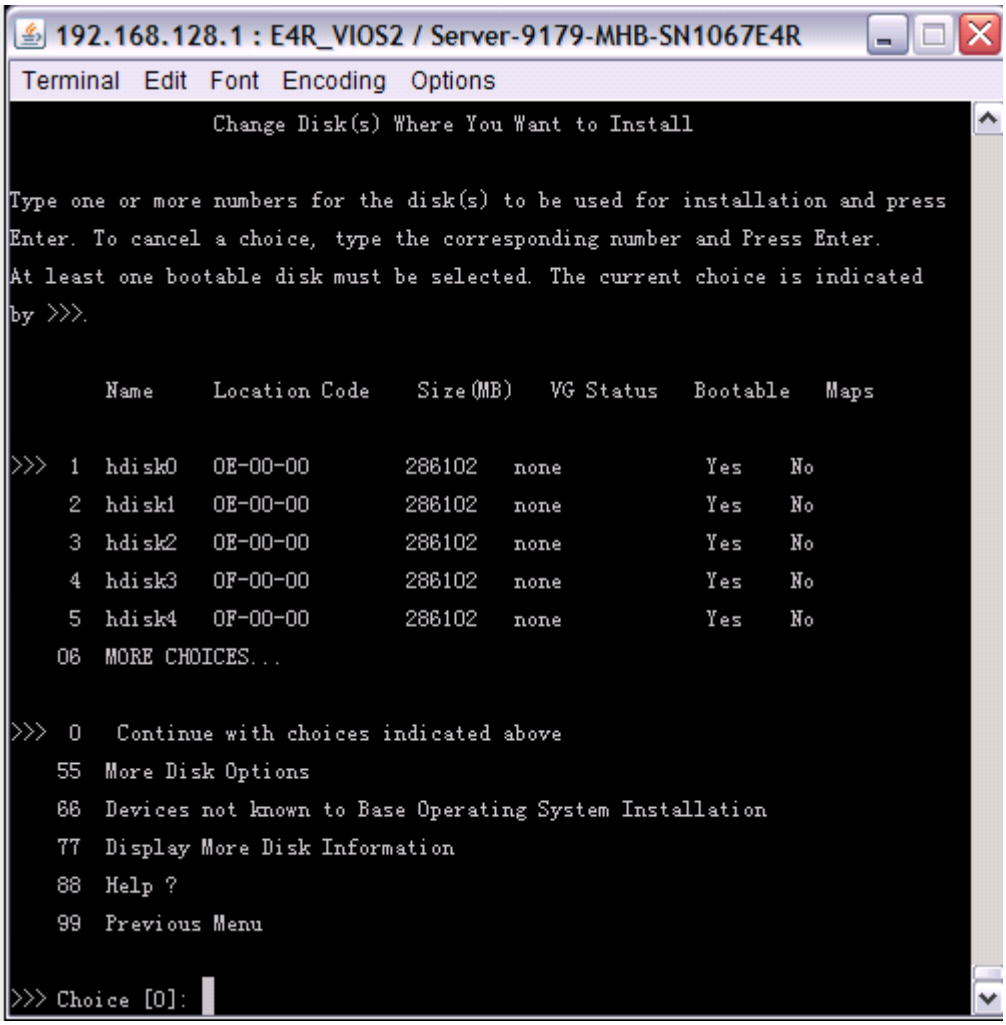
按 1 选择英语作为安装过程中使用的语言


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Welcome to Base Operating System
Installation and Maintenance
Type the number of your choice and press Enter. Choice is indicated by >>>.
>>> 1 Start Install Now with Default Settings
2 Change/Show Installation Settings and Install
3 Start Maintenance Mode for System Recovery
4 Configure Network Disks (iSCSI)
5 Select Storage Adapters
88 Help ?
99 Previous Menu
>>> Choice [1]: 2

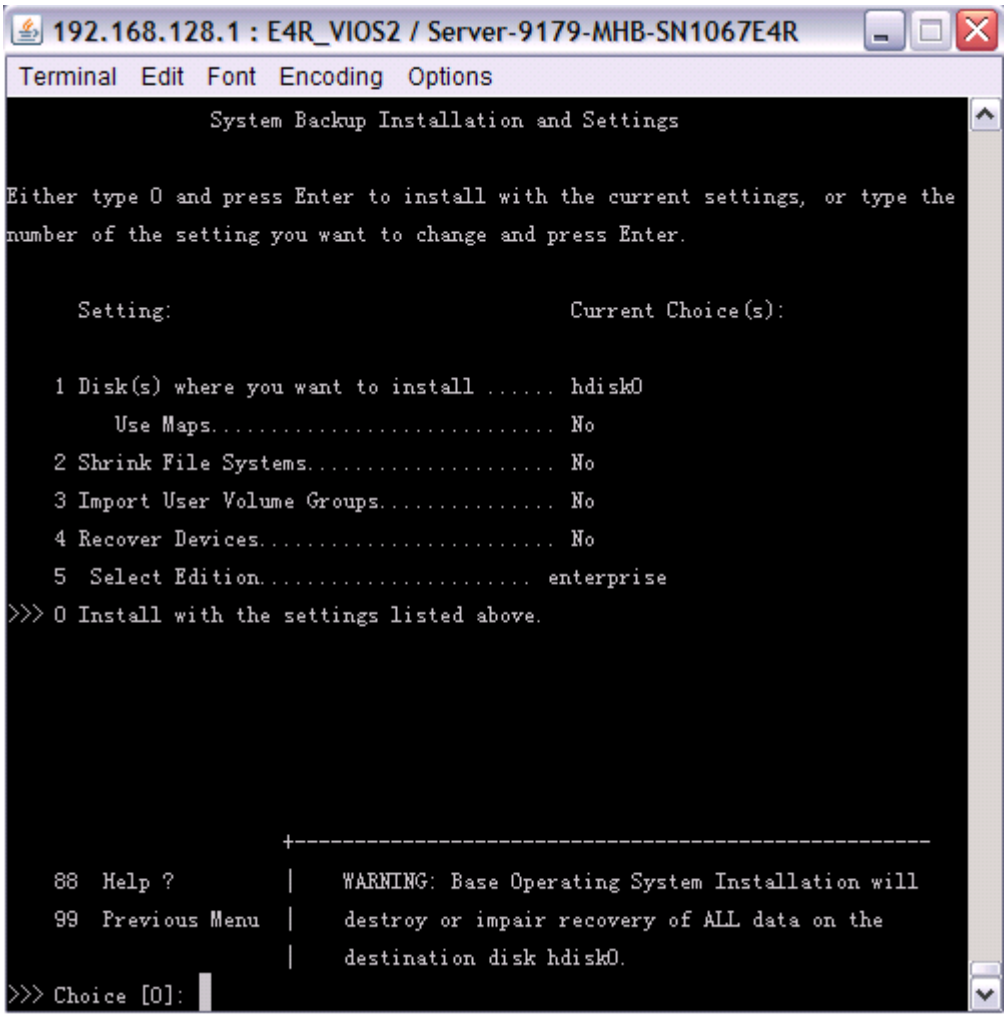
按 2 更改安装选项



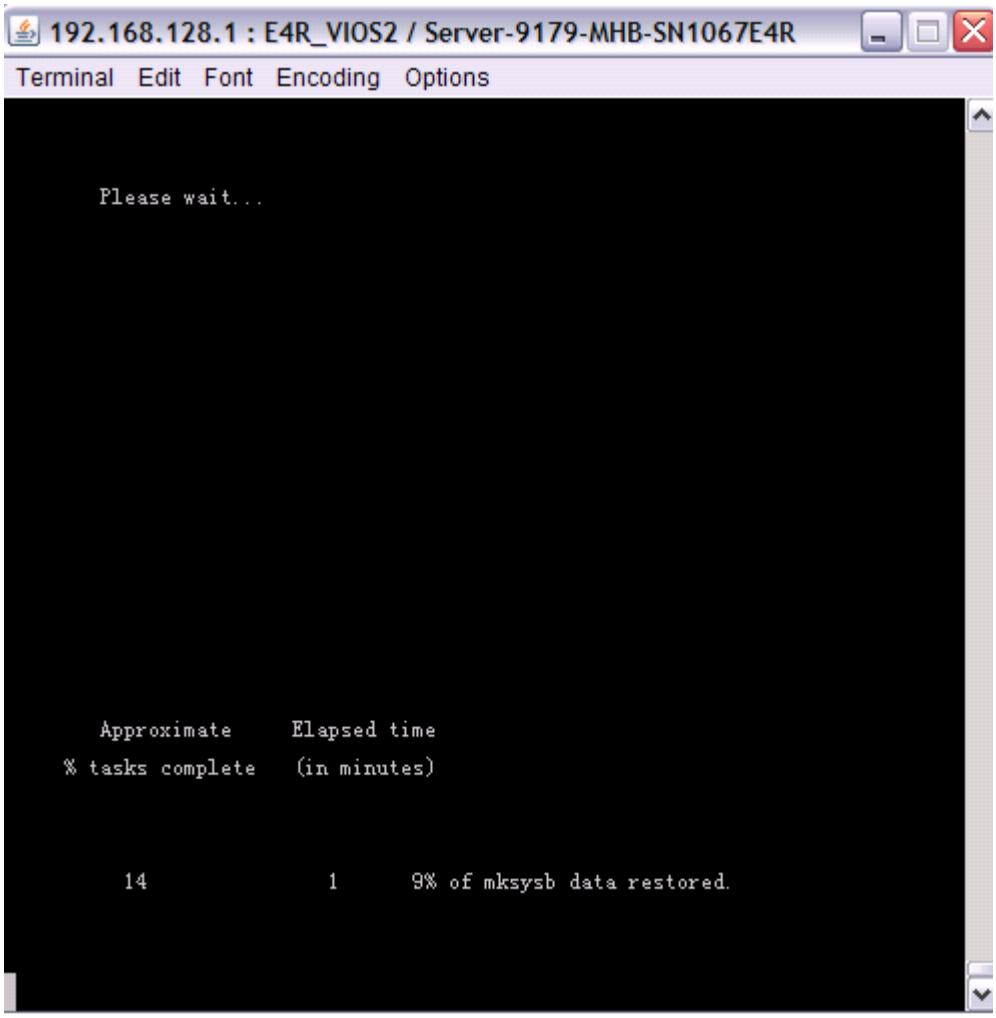
按 1 选择用于安装 VIOS 的磁盘



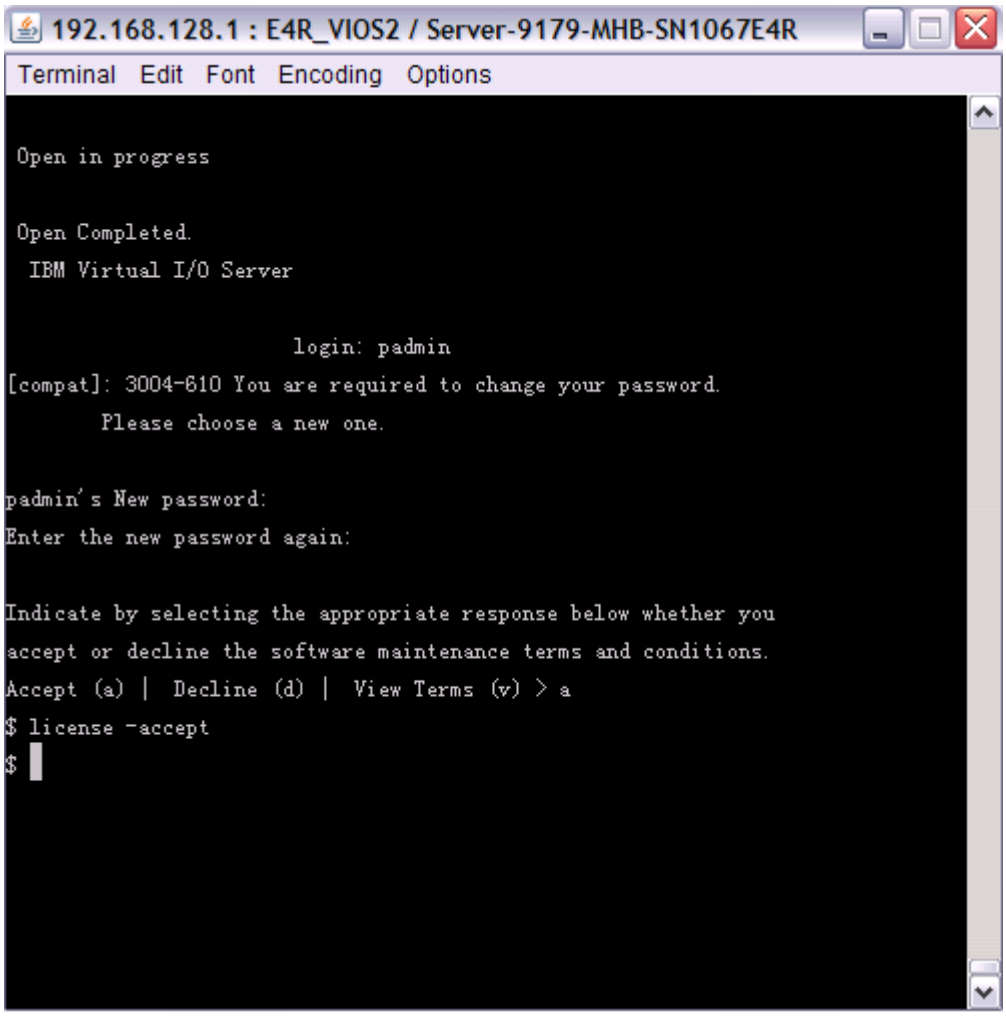
选择需要的磁盘，按 0 确认



按 5 选择安装版本为企业版，并按 0 开始安装


192.168.128.1 : E4R_VIOS2 / Server-9179-MHB-SN1067E4R
Terminal Edit Font Encoding Options
Please wait...
Approximate % tasks complete
Elapsed time (in minutes)
14 1 9% of mksysb data restored.

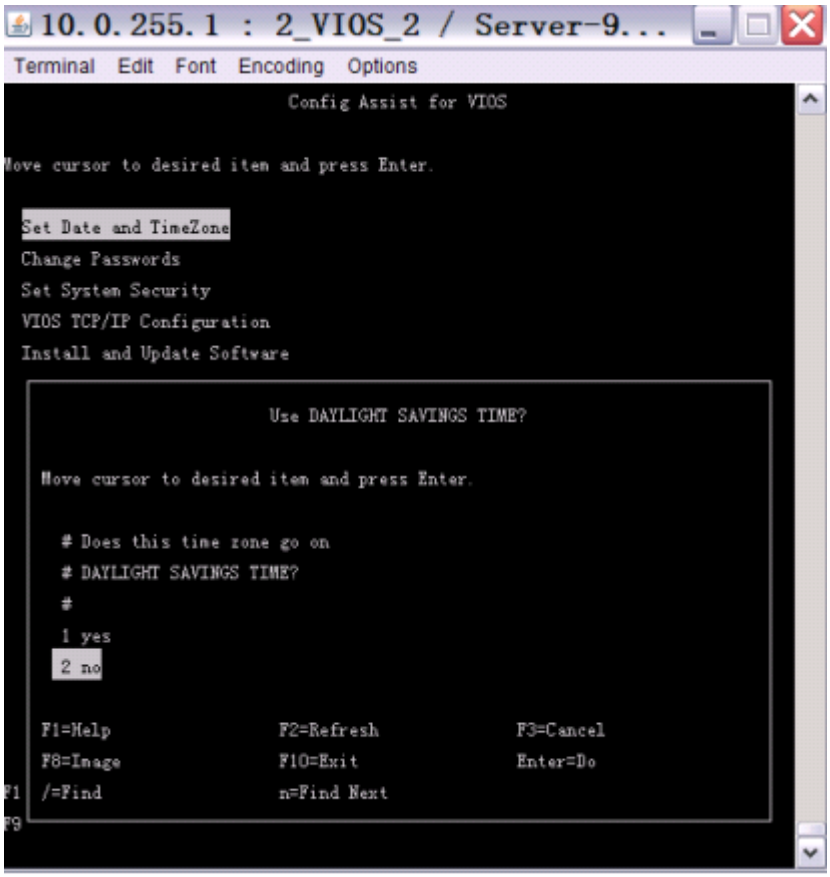
安装进行



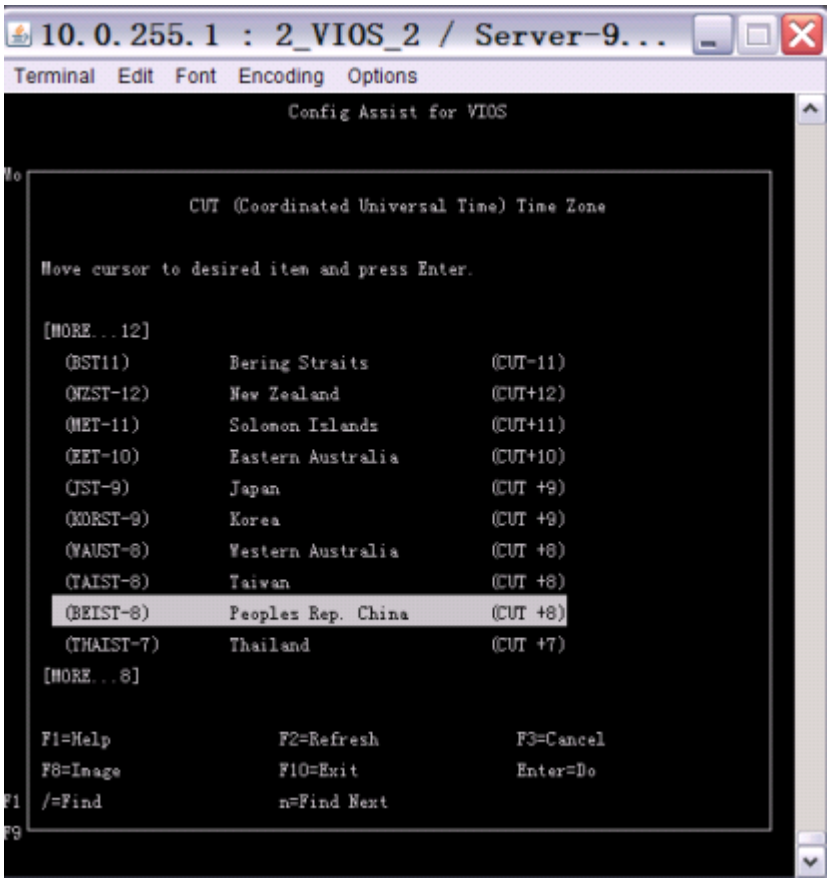
安装结束后，会自动重启，首先需要设置用户 padmin 的密码，然后输入 a 接受 license 进入系统，进入系统后，还需要输入 license –accept 才能进行后续的操作。

VIOS 安装完成后，需要设置一些基本参数，比如时区和时间，VIOS 下有类似 AIX 下 smitty 的工具叫 cfgasssit,可以做一些基本的配置，直接输入 cfgassist 即可:

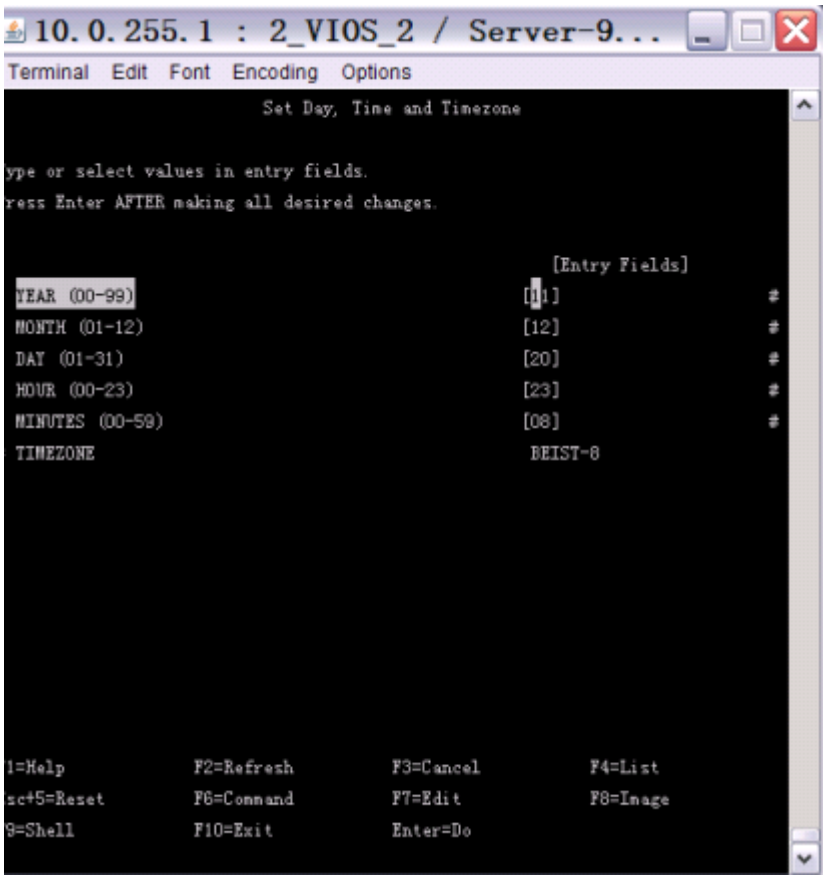
$cfgassist



选择”Set Date and TimeZone”，选择”2 no”不要夏令时


10.0.255.1 : 2_VIOS_2 / Server-9...
Terminal Edit Font Encoding Options
Config Assist for VIOS
CUT (Coordinated Universal Time) Time Zone
Move cursor to desired item and press Enter.
[MORE...12]
(BST11) Bering Straits (CUT-11)
(NZST-12) New Zealand (CUT+12)
(MET-11) Solomon Islands (CUT+11)
(EET-10) Eastern Australia (CUT+10)
(JST-9) Japan (CUT +9)
(KORST-9) Korea (CUT +9)
(WAUST-8) Western Australia (CUT +8)
(TAIST-8) Taiwan (CUT +8)
(BEIST-8) Peoples Rep. China (CUT +8)
(THAIST-7) Thailand (CUT +7)
[MORE...8]
F1=Help F2=Refresh F3=Cancel
F8=Image F10=Exit Enter=Do
/=Find n=Find Next

选择正确的时区，



更改日期和时间，完成了日期和时间的更改。

由于 VIOS 是一个定制的 AIX，默认的 padmin 账户权限很少，如果需要进入 VIOS 的 root 权限，输入 oem_setup_env 即可获取 VIOS 的 root 权限:

$oem_setup_env

3.2.2 VIOS 的镜像

如果 VIOS 安装在本地磁盘，建议做镜像进行保护，VIOS 也提供相应的命令进行操作，如下：

1) 加入另一块磁盘 hdisk1 到 rootvg 中

$extendvg rootvg hdisk1

2) 镜像 rootvg

$mirrorios –f hdisk1

镜像完成自动启动

3) 重建 boot image，切换到 root 权限下

#oem_setup_env

#bosboot –a –d /dev/hdisk0

#bosboot –a –d /dev/hdisk1

4) 设置启动列表

#bootlist –o -m normal hdisk0 hdisk1

5) 查看启动列表

$bootlist –mode normal -ls

3.3 VIOS 升级

升级 VIOS 需要登陆 IBM 官方网站下载升级补丁：

http://www-933.ibm.com/support/fixcentral/

每个升级包都有对应的 README，需要认真阅读，升级的基本步骤如下：

在升级 VIOS 之前需要执行以下命令来 commit 以前没有 commit 的软件包:

$updateios –commit

在 IBM 官方网站下载 VIOS 的补丁，通过 ftp 传到 VIOS 的目录下，通过以下命令来执行升级：

$ updateios –install –accept –dev /patch_directory

不同的升级包升级方法可能有差异，具体参考 readme。

升级完成执行 shutdown –restart 命令来重启 VIOS，然后通过 ioslevel 命令来确认升级后的版本信息。

对于单一 VIOS 配置的环境，如果要进行 VIOS 的升级，需要关闭全部 VIOC 分区。对于冗余 VIOS 配置的环境，可以通过轮流升级的方式来保证 VIOC 分区运行的连续性。

3.4 VIOS 管理

本章节主要介绍 VIOS 的一些日常管理，比如备份、恢复，还有虚拟媒体库的建立，DLPAR 的操作等

3.4.1 VIOS 的备份与恢复

3.4.1.1 VIOS 备份

对于 VIOS 的备份，通常如下一些组件发生变化之前，建议备份:

- 外部设备的配置，比如 SAN 交换机和存储
- 内存、CPU、虚拟和物理设备的变化
- VIOS 操作系统版本的变化
- 自定义虚拟设备和物理设备对应关系的变化

VIOS 的备份仅包含 VIOS 自身的操作系统和数据，不包括客户端分区的备份，如果客户端分区需要备份，遵循客户端分区备份的原则。

VIOS 的备份有三种主要的方式，通过命令 backupios 来进行：

- 备份到磁带
- 备份到 DVD-RAM
- 备份到远程的一个文件

备份到远程为一个 tar 文件是最常见的方式，如下示例，备份到一个远程的 NFS 目录

```
$ mount nim:/export/vios_backup /mnt
$ backupios -file /mnt -nomedialib
Backup in progress.  This command can take a considerable amount of
time to complete, please be patient...
```

其中 nomedialib 参数表示部分 VIOS 媒体库里面的内容，减少备份的大小，backupios 会创建一个名称为 nim_resources.tar 的文件，包含了所有能恢复 VIOS 的文件，包括 VIOS 的 mksysb，bosinst,data 文件，网络引导映像，还有 SPOT 引导文件，有了这些

文件，就可以从 NIM 或者 HMC 来恢复 VIOS 系统。

viosbr 命令：

viosbr 这个命令主要备份 VIOS 的虚拟和逻辑配置，在 VIOS 变更之前，即使不执行 VIOS 的全备份，建议至少通过 viosbr 备份一下以便恢复，viosbr 可以设置计划任务定期执行，以下设置每天执行，保留 7 个文件：

```
$ viosbr -backup -file vios22viosbr -frequency daily -numfiles 7
Backup of this node (vios22) successful
$ viosbr -view -list
vios22viosbr.01.tar.gz
```

备份用户自定义设备：

viosbr 已经备份了 VIOS 的虚拟和逻辑配置，可以通过如下命令查看：

```
$ viosbr -view -file vios22viosbr.01.tar.gz -mapping

Details in: vios22viosbr.01
SVSA                Physloc                          Client Partition ID
------------------ -------------------------------- --------------------
vhost0              U8233.E8B.061AB2P-V2-C30         0x00000003

VTD                   rootvg_lpar01
Status                Available
LUN                   0x8100000000000000
Backing Device        hdisk3
Physloc               U78A0.001.DNWHZS4-P2-D6
Mirrored              false

SVEA    Physloc
------- --------------------------------------
ent6    U8233.E8B.061AB2P-V2-C111-T1

VTD                   ent11
Status                Available
Backing Device        ent10
Physloc               U78A0.001.DNWHZS4-P1-C6-T2
```

这些备份的配置对于恢复到同一个机器同一个分区，相应的 mapping 关系可以恢复，但是如果恢复到不同的机器或者不同的分区，由于对应的物理设备发生了变化，无法自动恢复，需要手工再建立 mapping 关系。所以仍然建议通过手工备份一些配置信息，比

如可以把相应的 mapping 关系手工记录在 EXCEL 文件中，便于查阅和维护。

总结：对于 VIOS 的备份，可以定期通过 viosbr 来备份 VIOS 的对应的逻辑和虚拟配置，在重要变更之前通过 backupios 备份(包含 viosbr 的备份信息)VIOS 到远程的服务器上，这样才能保证一份完整的备份。同时建议为 VIOS 中自定义的虚拟设备，虚拟设备和物理设备对应的 mapping 建立一份维护文件比如 EXCEL 表格，便于查阅和维护。

3.4.1.2 VIOS 恢复

恢复 VIOS 通常通过 HMC 和 NIM 来进行，

通过 HMC 恢复：

假设通过 backupios 备份的 nim_resource.tar 放在一个远程目录，并且此目录已经通过 NFS export 出来，在 HMC 下执行 installios，输入相应的参数：

```
hscroot@hmc9:~> installios -p vios22 -i 172.16.22.33 -S 255.255.252.0
-g 172.16.20.1 -d 172.16.20.41:/export/vios_backup/ -s
POWER7_2-SN061AB2P -m 00:21:5E:AA:81:21 -r default -n -P auto -D auto
...
...Output truncated
...
# Connecting to vios22
# Connected
# Checking for power off.
# Power off complete.
# Power on vios22 to Open Firmware.
# Power on complete.
# Client IP address is 172.16.22.33.
# Server IP address is 172.16.20.111.
# Gateway IP address is 172.16.20.1.
# Subnetmask IP address is 255.255.252.0.
# Getting adapter location codes.
# /lhea@200000000000000/ethernet@200000000000002 ping successful.
# Network booting install adapter.
# bootp sent over network.
# Network boot proceeding, lpar_netboot is exiting.
# Finished.
```

通过 NIM 恢复：

由于通过 backupios 备份的文件是一个 tar 文件，而通过 NIM 恢复 VIOS 需要一个 mksysb

文件，所以需要把这个 tar 文件解压找出 mksysb 文件，找到相应的 mksysb 文件后，在 NIM 服务器上定义 NIM 的资源，如下:

```
# nim -o define -t mksysb \
-a server=master \
-a location=/export/vios_backup/vios22.mksysb vios22_mksysb
# nim -o define -t spot \
-a server=master \
-a location=/export/vios_backup/spot \
-a source=vios22_mksysb vios22_spot
# nim -o bos_inst \
-a source=mksysb \
-a mksysb=vios22_mksysb \
-a spot=vios22_spot \
-a installp_flags=-agX \
-a no_nim_client=yes \
-a boot_client=no \
-a accept_licenses=yes vios22
```

定义好资源后，通过 NIM 恢复 VIOS 与恢复 AIX 操作一样。

3.4.2 VIOS 的 DLPAR 操作

通过 DLPAR 操作更改 CPU、内存、IO 适配器配置不会同时更改到分区的概要文件中，这样分区关闭再启动就会丢失 DLPAR 操作后的配置，一个好的习惯是在进行 DLPAR 之前先更改概要文件中的配置，再执行 DLPAR 操作，这样可以保证分区的概要文件和当前运行的配置一致。

另一种方式是执行了 DLPAR 操作之后，把当前运行的配置保存为概要文件，如下：

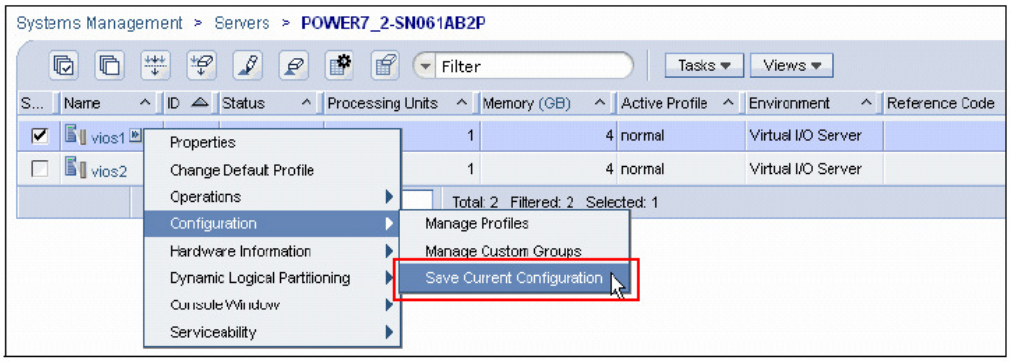
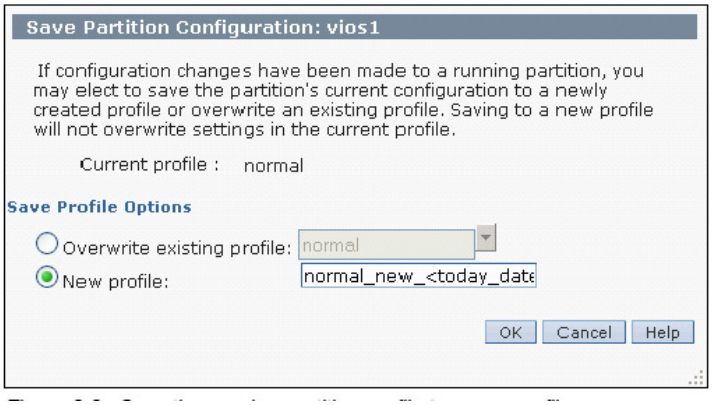


Figure 3-2 Partition context menu



可以选择新定义一个文件，也可以选择覆盖当前的概要文件(Overwrite existing profile 方式似乎更好，避免多个 profile)。

3.4.2.1 VIOC 增加虚拟光纤卡 VFC(Virtual Fibre Channel Adapters)

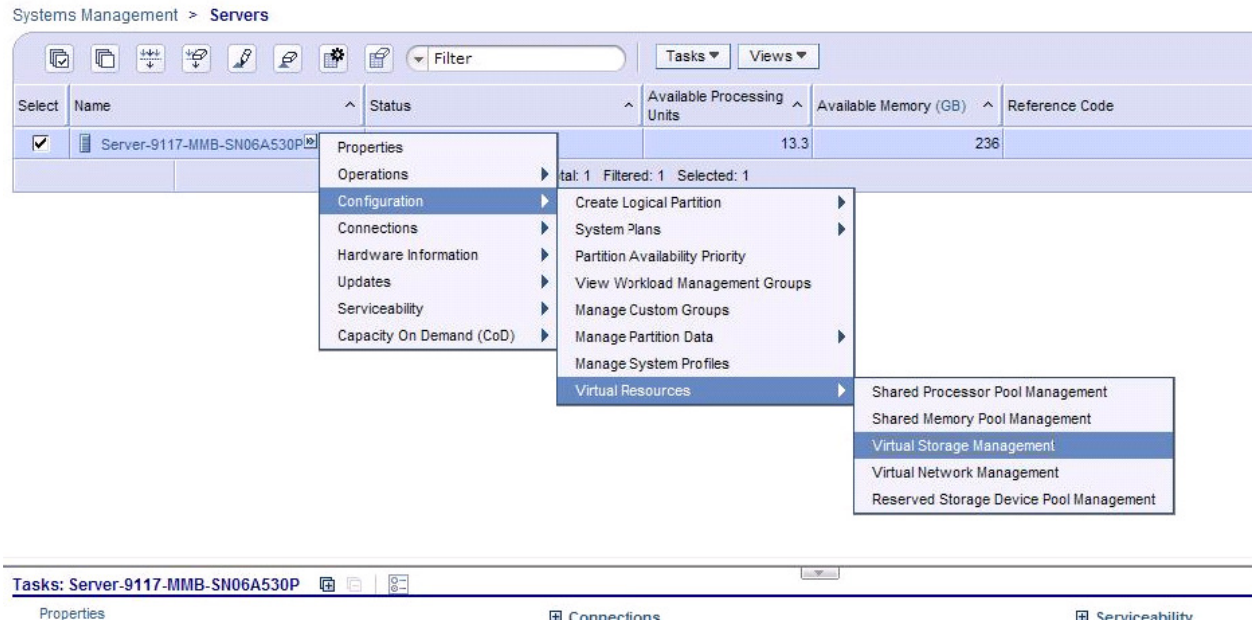
VFC 的 wwpn 号是在 VIOC 上创建 VFC 时才进行分配，VIOS 上创建 VFC 时并不会分配 wwpn 号，所以这里是否改成 VIOC 动态增加虚拟光纤卡？

VFC 的 wwpn 号是通过 Power 服务器的 Hypervisor 分配的，每次分配都是唯一的，已分配的 wwpn 号在 VFC 卡删除后并不会被回收，一旦 wwpn 用完需要向 IBM 额外申请。如果通过 DLPAR 增加一个 VFC，分配了一对 wwpn，这时再通过更改 profile 的方式增加一个 VFC，又会分配一对新的 wwpn，这样 wwpn 就不一致了，因此如果通过 DLPAR 来增加虚拟光纤卡，只能选择 SAVE current profile 的 Overwirte existing profile 方式来保持 DLPAR 操作后与 profile 的一致。

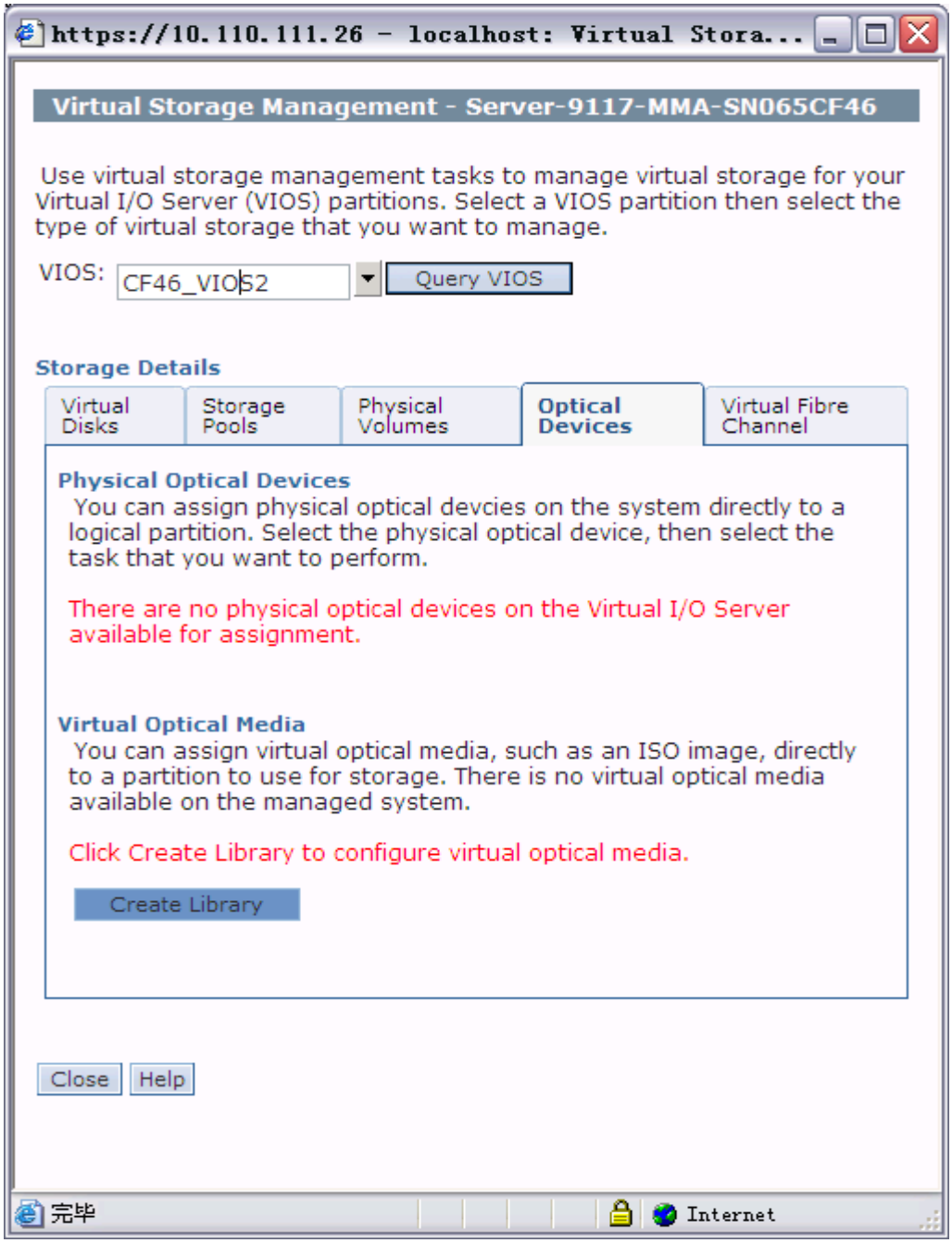
3.4.3 VIOS 的虚拟媒体库创建

VIOS 有一个重要的功能就是可以把 ISO 镜像做成虚拟光驱，映射给客户端分区使用，还可以同时使用，这样的好处是可以不用服务器自带的物理光驱也可以安装客户端分区的操作系统，在 VIOS 上创建媒体库步骤如下：

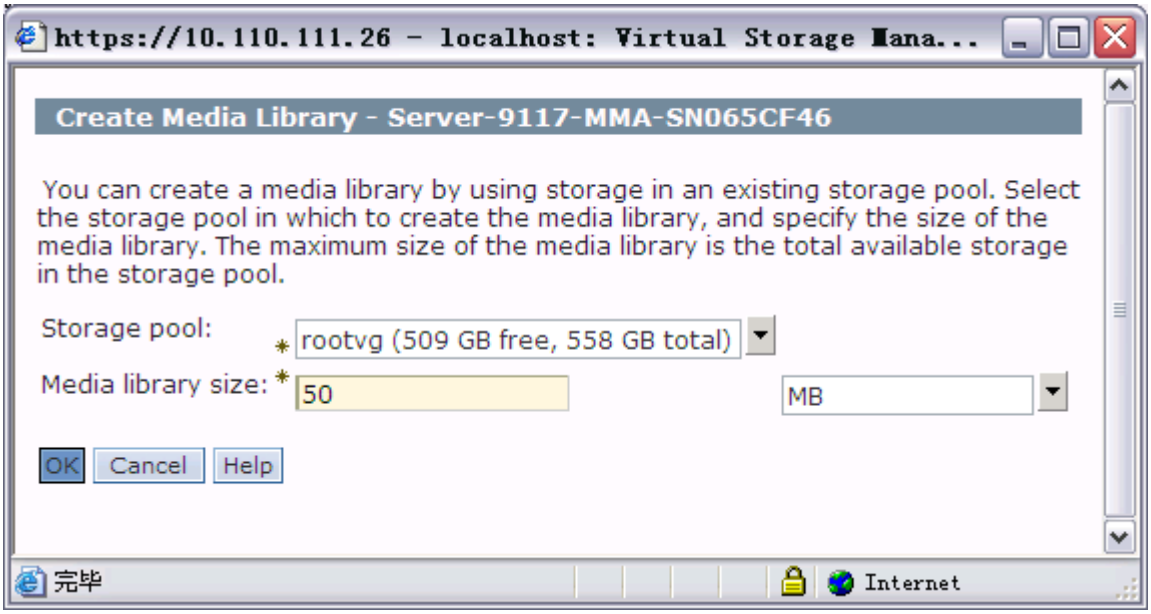
在 HMC 中选中物理服务器->配置->虚拟资源->虚拟存储管理



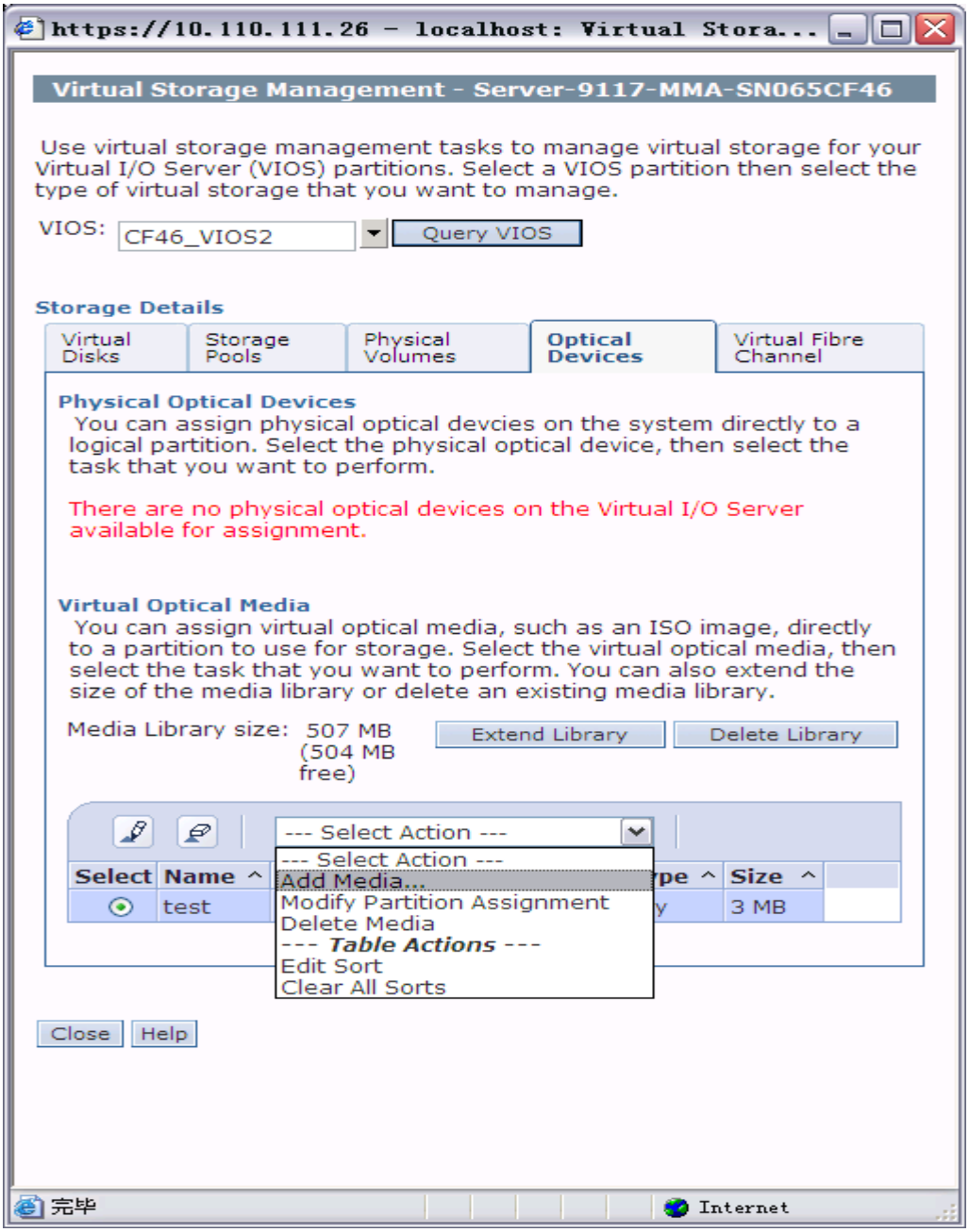
选择创建虚拟媒体库的 VIOS，然后点击 Query VIOS，然后选择 optical devices，点击 Create Library



填入需要创建的媒体库的大小，根据需要，输入合适的值



创建虚拟光驱，从下拉菜单中选择添加介质文件

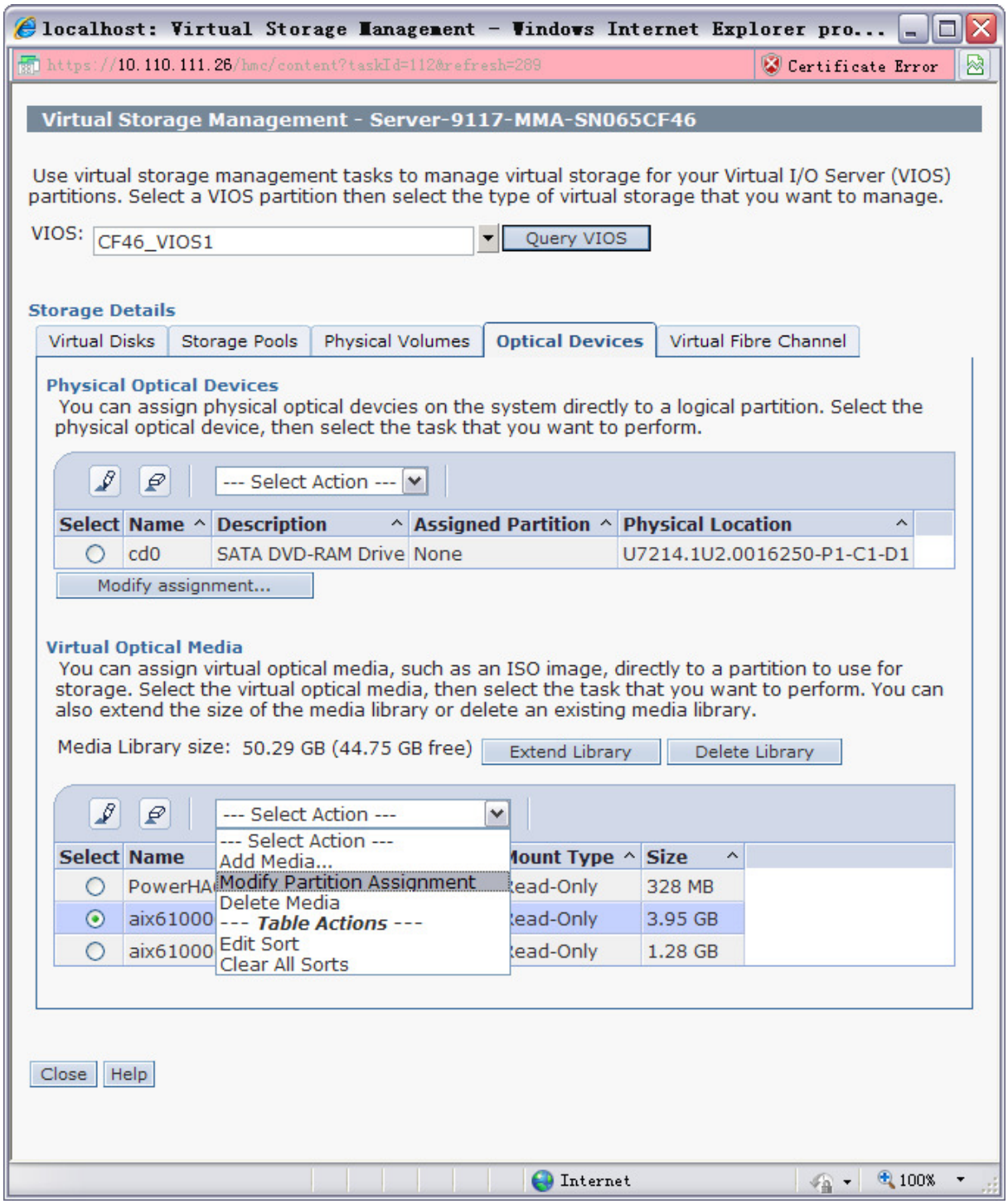

https://10.110.111.26 - localhost: Virtual Stora...
Virtual Storage Management - Server-9117-MMA-SN065CF46
Use virtual storage management tasks to manage virtual storage for your Virtual I/O Server (VIOS) partitions. Select a VIOS partition then select the type of virtual storage that you want to manage.
VIOS: CF46_VIOS2
Query VIOS
Storage Details
Virtual Disks
Storage Pools
Physical Volumes
Optical Devices
Virtual Fibre Channel
Physical Optical Devices
You can assign physical optical devcies on the system directly to a logical partition. Select the physical optical device, then select the task that you want to perform.
There are no physical optical devices on the Virtual I/O Server available for assignment.
Virtual Optical Media
You can assign virtual optical media, such as an ISO image, directly to a partition to use for storage. Select the virtual optical media, then select the task that you want to perform. You can also extend the size of the media library or delete an existing media library.
Media Library size: 507 MB (504 MB free)
Extend Library
Delete Library
--- Select Action ---
--- Select Action ---
Add Media...
Modify Partition Assignment
Delete Media
--- Table Actions ---
Edit Sort
Clear All Sorts
Select
Name
Size
test
3 MB
Close
Help
完毕
Internet

选择导入的方式，从光驱导入还是从现有的文件添加

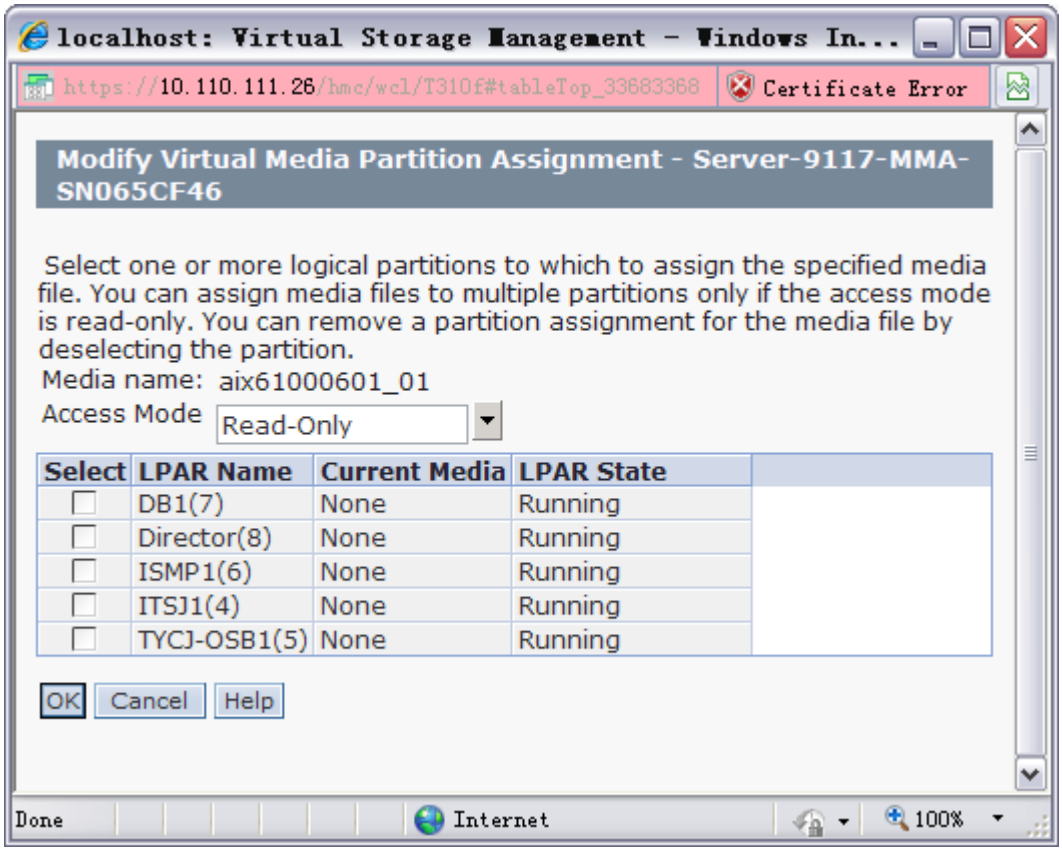
这里以从光驱导入为例，点击 OK

文件添加结束后，虚拟光驱就建成了，可以分配给客户端使用，如下：

选择要挂载的介质文件名称，在下拉菜单中选择修改分区分配。

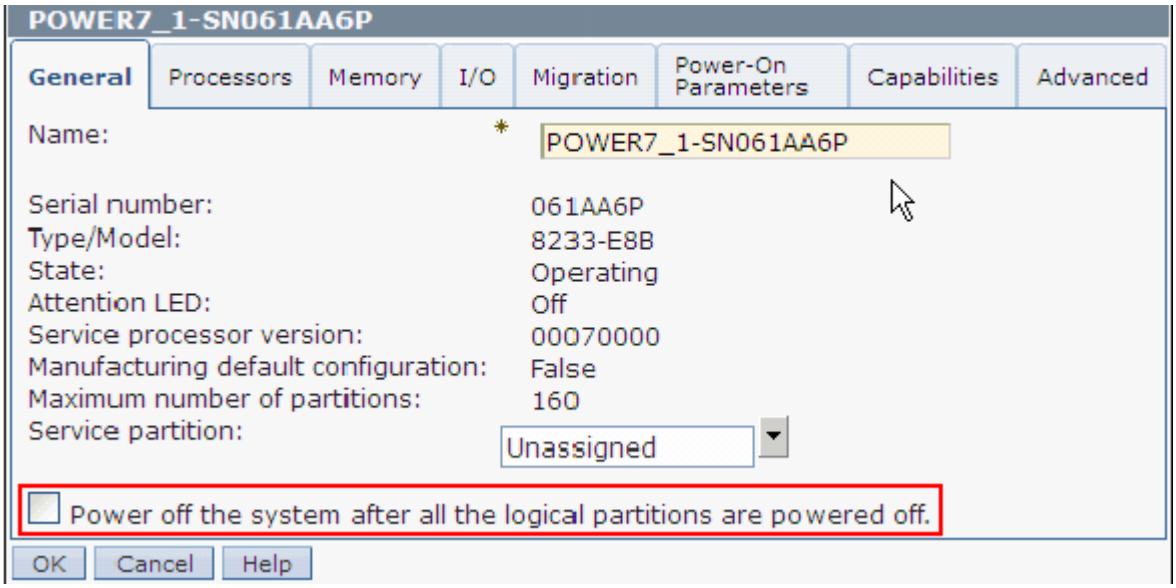


在弹出的菜单中选择要分配给哪个分区，选择后点击 OK。

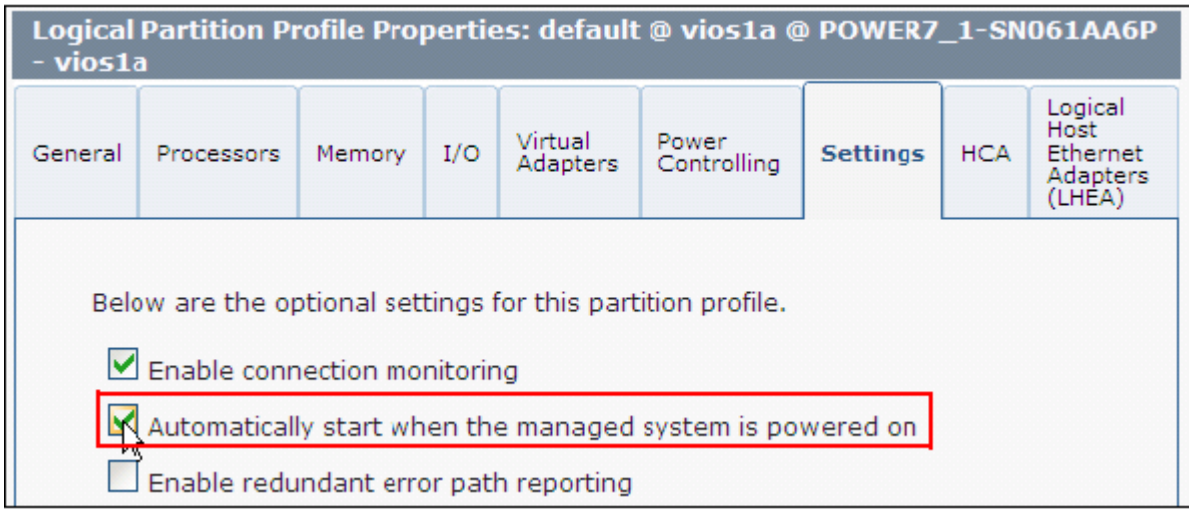


3.4.4 服务器的关闭与启动

对于实施了 PowerVM 的服务器，建议所有的分区启动或关闭都由手动来进行，而不是随服务器启动而启动，关闭而关闭，



比如上图选项，服务器属性里面不要勾上所有分区关闭后，服务器也关闭



对于分区属性，也不要选择随服务器启动而启动

对于分区启动的顺序和原则如下：

1. VIOS

2. CPU 或者内存资源较大的 VIOC 分区

3.其他 VIOC 分区

对于分区关闭的顺序如下：

1. VIOC 分区

2. VIOS 分区

4 VIOS 网络配置

Power 服务器提供了丰富的虚拟化网络配置，下面主要介绍常见的一些配置

4.1 基本网络参数考虑

常用的网络术语：

术语	解释
Link aggregation(LA)	是指多块物理网卡聚合在一起作为一个网卡使用，增加网络带宽，提供高可用性，目前有两种常用的配置模式：EtherChannel 和 802.3ad，
Network interface Backup(NIB)	NIB 是提供了一个备份网卡，但是平常不激活，在主网卡失效的情况下，接管工作；主网卡和备份网卡接在不同的交换机上
VLAN tagging	标准的 802.1Q 协议，允许一个网络连接通过多个 VLAN，也叫 VLAN Trunking

在配置你的 PowerVM 网络之前，建议参考以下步骤逐步确认配置：

- 跟网络管理员沟通，保持术语的一致
- 记录下需要的信息
- 使用 VLAN Tagging 的话，保证虚拟网络和外部网络的 VLAN ID 一致
- 对于 VIOC 分区，使用 PVID，不要在一个虚拟网卡配置多个 VLAN，也不要使用 AIX 提供的 VLAN tagging 功能，保证配置简单

- 对于 VIOS 来说，尽量使用可以热插拔的物理网卡
- 在允许的情况下，使用双 VIOS 保证冗余
- VIOS 上的物理 IO 卡尽量分布在不同的柜子上
- 使用 HMC 的 vterm 在 VIOS 上配置网络
- 3358 的 VLAN ID 不可用，服务器保留
- 每个虚拟网卡支持的最大 VLAN 数是 21 个，包括 20 个 VLAN ID，1 个 PVID
- 每个 SEA 支持最多 16 个虚拟网卡
- 物理网卡的聚合配置最多支持 8 主 1 备，8 主在一个物理交换机，1 备在另一个物理交换机
- 虚拟网卡的最大帧为 65408 bytes
- 如果是 SEA Failover 的配置，必须为每对 SEA 配置一个心跳网络的控制通道，即 Control Channel，Control Channel 所在虚拟网卡的 PVID 必须在两个 VIOS 上设置为一致，此虚拟网卡无需设置为可以访问外部网络
- 如果是 SEA Failover 的配置，在一个 VIOS 需要设置成 SEA 的所有虚拟网卡的优先级要是一致的，不能有不同。

4.1.1 SEA(Shared Ethernet Adapter)考虑

以下是创建 SEA 之前的考量点：

无 VLAN tagging：

- 一个虚拟网卡一个 VLAN，一个虚拟网卡对应一个 SEA
- 一个虚拟网卡一个 VLAN，多个虚拟网卡对应一个 SEA

VLAN tagging：

- 一个虚拟网卡对应所有的 VLAN，一个虚拟网卡对应一个 SEA

- 一个虚拟网卡对应一个 VLAN，多个虚拟网卡对应一个 SEA
- 以上两种的混合

对于满足如下条件的 PowerVM 环境，可以动态在线增减 VLAN：

VIOS 2.2.0.0 以上

Power7 服务器

服务器微码：AH720_064+，AM720_064+，AL720_064+以上

HMC：V7.7.2.0 MH01235 以上

在设置你的虚拟化网络之前，建议列出一个表格，把相应的物理，虚拟网卡的配置列出来，以便做配置，表格可以如下表，也可以自己定义一个适合的格式：

	Physical Ethernet	***Link Aggregation***	***Virtual Ethernet***	***SEA***	***Trunk Priority***	***PVID***	***VLAN (if VLAN tagging)***
VIOS	*ent0*	*ent3*	*ent4*	*ent7*	*1*	*100*	*202/702*
	ent1		*ent5*		*1*	*200*	*421/701*
	ent2		*ent6*		*control channel*	*99*	

4.1.2 网络参数修改

VIOS 上物理网卡的参数，修改建议：

jumbo_frames=yes, large_send=yes, large_recieve=yes, flow_ctrl=yes

VIOS 上 SEA 参数的建议：

jumbo_frames=yes, large_send=yes, large_recieve=yes

注：如果客户端分区是 LINUX 或者 IBM i，以上的 large_send 和 large_receive 不建议

启用，会降低性能

4.2 单 VIOS 网络

对于单 VIOS 网络，物理网卡建议多块聚合，以两块物理网卡 ent0 和 en1 聚合为 ent2，虚拟网卡 ent3，PVID 为 100，VLAN ID 为 200,和 300 为例，配置 SEA 的命令如下：

$mkvdev –sea ent2 –vadapter ent3 –default ent3 –defaulid 100 largesend=1 large_receive=yes

命令执行完毕后，会生成一个新的 ent 设备，可以通过如下命令来查看配置：

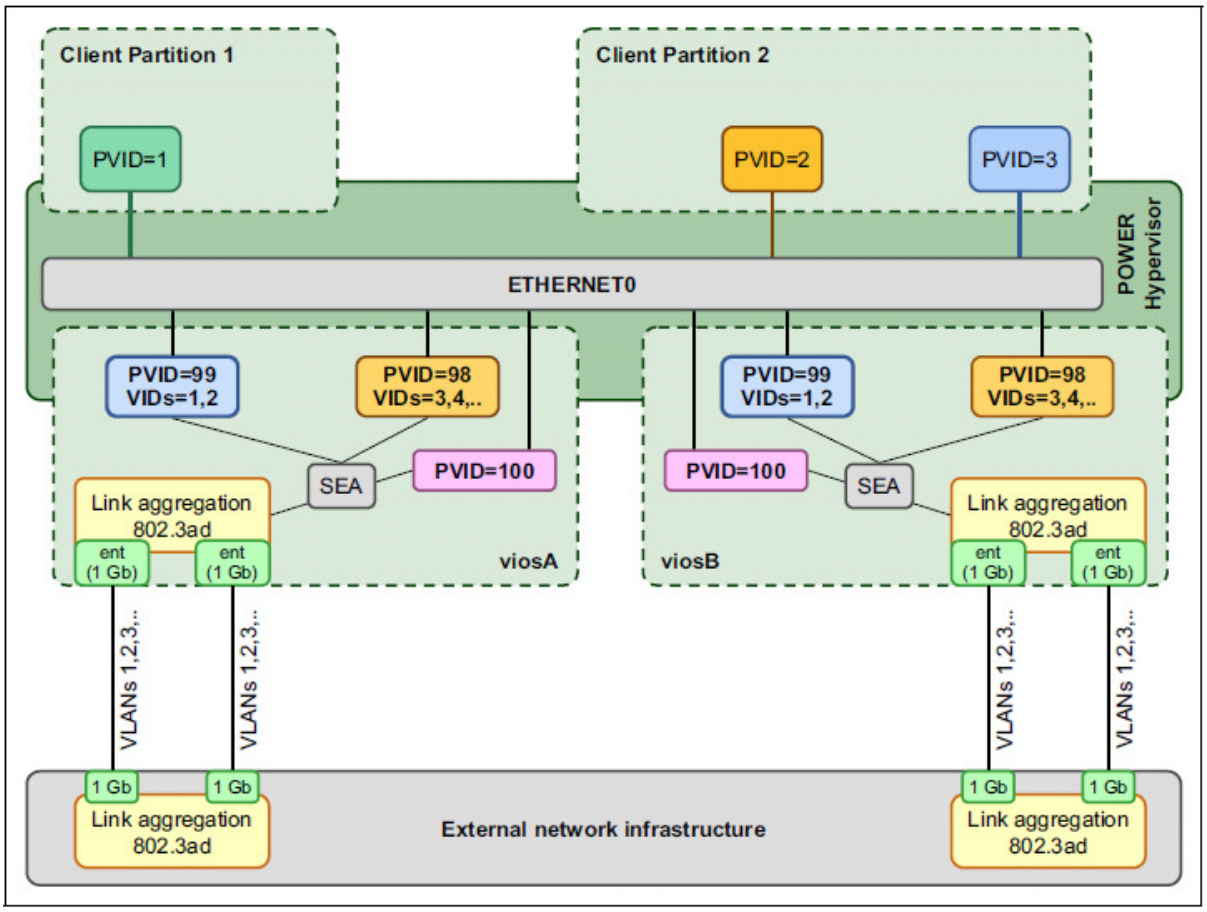
```
$ lsmap -all -net
SVEA   Physloc
------ --------------------------------------------
ent2   U9117.MMA.101F170-V1-C11-T1

SEA                      ent3
Backing device           ent0
Status                   Available
Physloc                  U789D.001.DQDYKYW-P1-C4-T1
```

4.3 双 VIOS 环境下虚拟网络的冗余

SEA 在双 VIOS 环境下的冗余主要有两种办法，一种是通过客户端的 NIB 方式，还有一种是通过 VIOS 端的 SEA failover，由于通过 VIOS 端的 SEA failover 配置更简单，推荐虚拟网络的冗余使用 SEA failover 这种方式。

4.3.1 SEA(Shared Ethernet Adapter)的冗余配置

随着 PowerVM 技术的不断更新，SEA failover 出了新的技术称为 load sharing,相比以前 failover 的好处是两个 VIOS 的 SEA 可以同时活动。



如上午所示，两个 VIOS 各有两块物理网卡聚合，每个 VIOS 各有三块虚拟网卡，一块是 PVID 为 99，VLAN ID 为 1,2；第二块是 PVID 为 98，VLAN ID 为 3,4；第三块 PVID 为 99，无 VLAN tagging。其中前两块为数据网卡，第三块为 control channel 控制通道使用的网卡，控制两个 VIOS 上 SEA 的 failover，上面配置的好处就是 VLAN 为 99,1,2 的网络走第一个 VIOS，VLAN 为 98,3,4 的网络走第二个 VIOS,这样不会造成其中一个 VIOS 闲置，造成物理网卡的浪费，尤其是针对 10Gb 网卡的环境。

假设物理网卡为 ent0 和 ent1，聚合后为 ent2，PVID 为 99 的网卡为 ent3,PVID 为 98 的网卡为 ent4，PVID 为 100 的网卡为 ent5，那么在两个 VIOS 上配置的命令分别如下：

viosA:

$mkvdev -sea ent2 -vadapter ent3 ent4 -default ent3 defaultid 99 -attr ha_mode=sharing ctl_chan=ent5 largesend=1 large_receive=yes

viosB:

$mkvdev -sea ent2 -vadapter ent3 ent4 -default ent4 defaultid 98 -attr ha_mode=sharing ctl_chan=ent5 largesend=1 large_receive=yes

关于 SEA Failover with Loadsharing 总结如下：

- 做 SEA 的虚拟网卡至少要两块，且在每个 VIOS 上的 priority 要一致
- 无 VLAN tagging 的示例：两个虚拟网卡一个 PVID 100，一个 PVID 200，那么一个走 VIOS1，另一个走 VIOS2
- VLAN tagging 的示例：两个虚拟网卡，一个 PVID 100，additional vlan 101 102 103；一个 PVID 200 additional vlan 202 203，那么，100、101、102、103 会走一个 VIOS；200、202、203 会走一个 VIOS，
- 做不到真正意义上的平均，只能保证做 SEA 的多块虚拟网卡均分在两个 VIOS 上，而每个虚拟网卡上的所有 VLAN 是一定在同一个 VIOS 上的

4.4 SEA 状态查看

SEA 配置完毕后，可以通过以下命令查看 SEA 的配置：

$lsmap –all net

```
$ lsmap -all -net
SVEA   Physloc
------ --------------------------------------------
ent2   U9117.MMA.101F170-V1-C11-T1

SEA                  ent3
Backing device       ent0
Status               Available
Physloc              U789D.001.DQDYKYW-P1-C4-T1
```

查看 SEA 的状态，可以通过 entstat –all entX 来查看

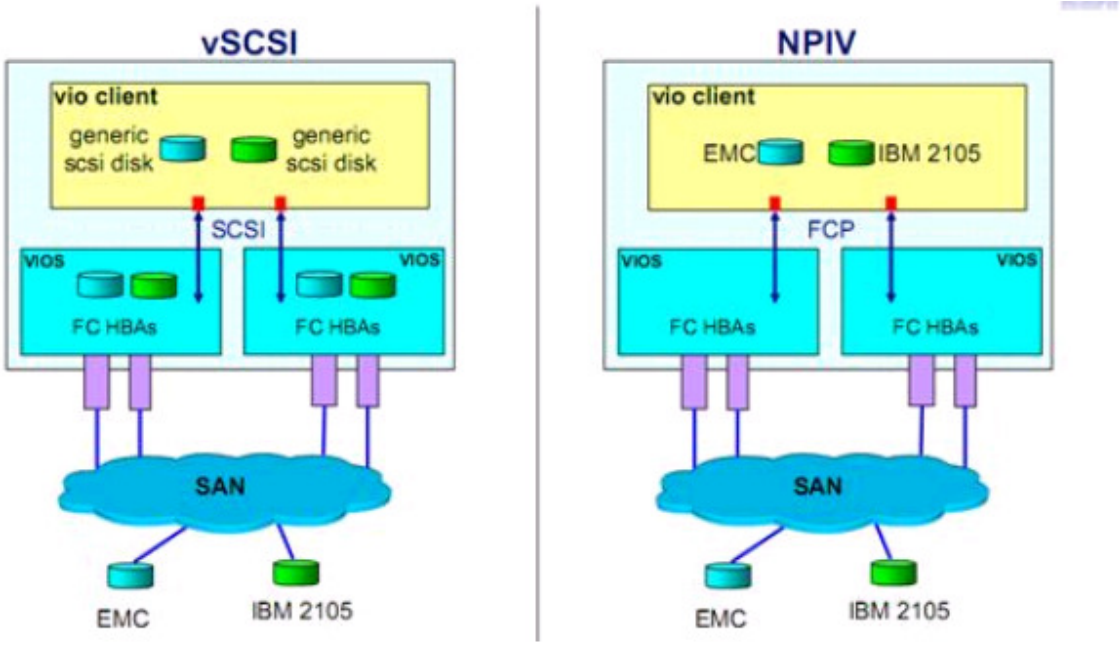
$entstat –all ent3

5 VIOS 存储配置

本章节介绍 VIOS 中存储资源的配置

5.1 VIOS 的存储考虑

VIOS 可以通过两种方式给 VIOC 提供存储资源访问，一种是 Virtual SCSI，另一种是 NPIV 方式，具体细节不再介绍，以下为两种典型的拓扑:



对于共享存储资源的 vSCSI 模式来说，异构存储是由 VIOS 统一汇集成一个单一块存储池，然后再以通用 SCSI LUN 的形式分配给客户端的 LPAR 中。SCSI 模拟和 SCSI target 的扮演的工作都是由 VIOS 来执行的。与 VSCSI 不同的是，在 NPIV 的过程中，VIOS 所起的作用是根本不同的。VIOS 只提供功能便于适配器的共享，没有设备级抽象或模拟。而且不像 VSCSI 那样提供存储虚拟的功能 ,VIOS 只提供 NPIV 的一个捷径，以便客户端的 LPAR 通过 FCP 口连接到 SAN 网络。

vSCSI 的优点：VIOC 使用 vSCSI 磁盘与传统使用本地磁盘方法一致，架构简单，易

于排障；且在交换机上需要划分的 ZONE 较少，只需要划分 VIOS 与存储通信的 ZONE, 无需 VIOC 与存储通信的 ZONE。

NPIV 的优点：VIOS 上无需访问磁盘，手工工作较少；VIOC 直接访问存储，可以实现负载均衡；LPM 操作简单，无需额外操作，无需手工在目标机器上做任何操作。

vSCSI 与 NPIV 的对比如下：

IO虚拟化方式	优势
vSCSI	标准 / 通用的界面;
	实现以下存储虚拟化功能: 磁盘 / 光盘 / 磁带 设备共享 可以虚拟化 non-SAN 存储 (iSCSI, SAS, parallel SCSI, etc)
	利用服务器的VIO服务器创建、管理存储池
	敏捷性： 快速的 LPAR 部署
NPIV	简化 provisioning (与 vSCSI Classic相比)
	简化管理 (包括 LPM)
	虚拟服务器所占用的存储空间和性能状况都可以很容易跟踪
	可以共享光纤卡资源而无需更换现有的工具、方案
	可以支持磁带库设备
	可以实现负载均衡 (active/active) 和 异构的 multipathing
	支持SCSI-3 Persistent Reserve
	Physical<---->Virtual 设备兼容 (使用原厂驱动)

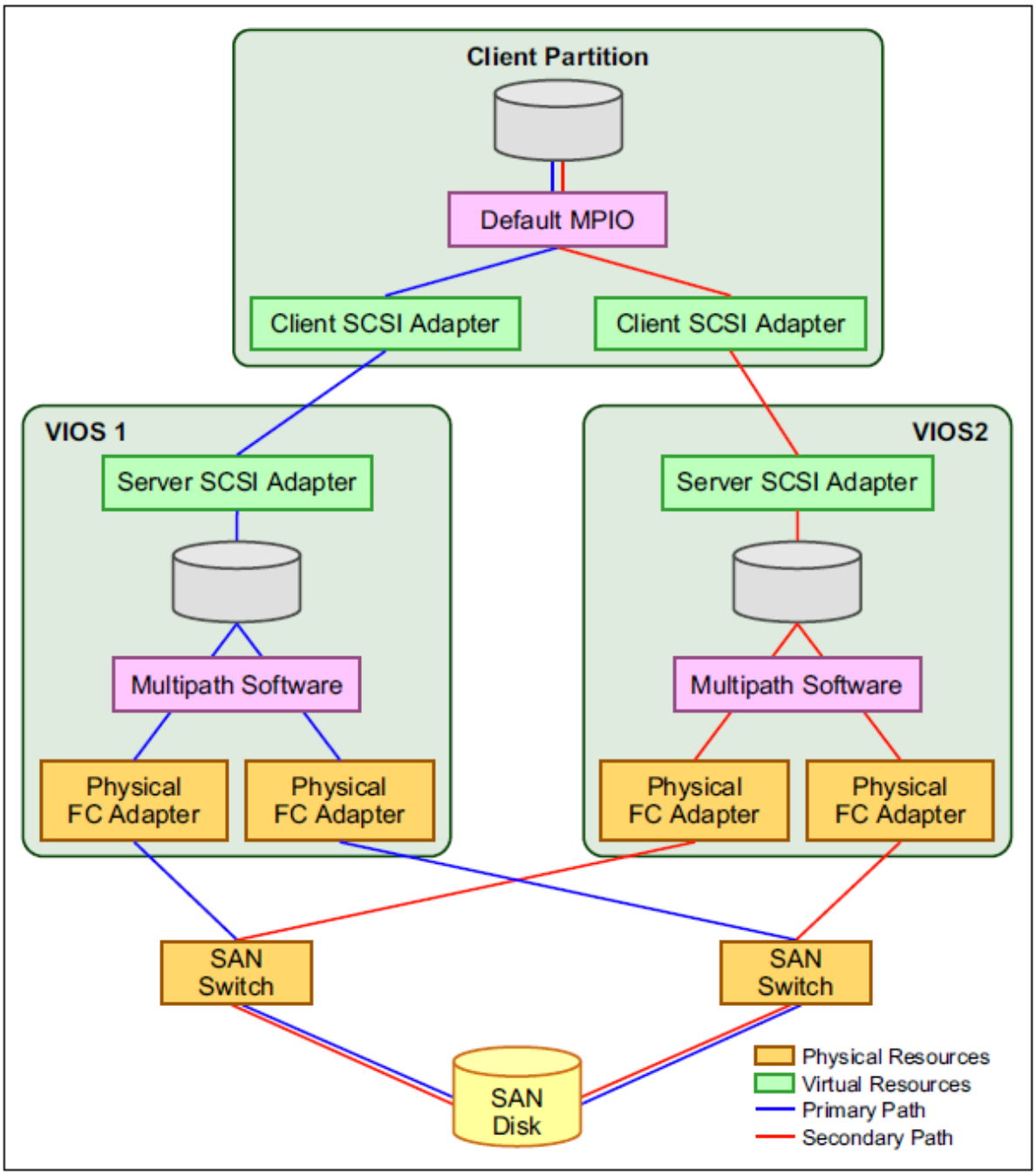
5.1.1 VIOS 的 rootvg 考虑

对于 VIOS 的 rootvg，建议安装在内置盘，通过内置盘镜像来实现冗余，镜像的方式在 3.2.2.已经有介绍，而 VIOC 的所有磁盘建议使用 SAN 存储，不管是通过 vSCSI 还是通过 NPIV 的方式。

5.1.2 多路径 Multipathing 考虑

Multipathing 多路径为服务器访问存储提供了负载均衡和冗余性，不管在 VIOS 还是 VIOC 都建议配置多路径，VIOS 上的多路径通过多块光纤卡实现；VIOC 不管是通过 VSCSI 方式还是 NPIV 方式，建议配置两个 VIOS，不同的时候如果是通过 VSCSI 方式，VIOC 不需要安装额外的软件，利用自带的 MPIO，且只能提供 Failover 方式，无 loadbalance 方式；如果是通过 NPIV 方式，VIOC 需要安装存储厂商提供的多路径软件，且根据存储的不同，有 failover 和 loadbalance 两种冗余方式。

以下是一个典型的 vSCSI 配置：



VIOS 通过两块光纤卡在 VIOS 层面实现多路径访问 SAN 存储，VIOC 通过 VSCSI 的 MPIO 访问 VIOS 上 SAN 存储分配的磁盘。

VIOS 支持的存储，可以参考如下链接：

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html#solutions

如果是非 IBM 的存储，建议阅读产品的说明，是否支持 VIOS。

5.1.3 Virtual SCSI 和 NPIV 的混合环境考虑

虽然 VSCSI 和 NPIV 技术不同，且各有优点，但是在同一个 VIOC 上可以同时使用两种技术，比如 VIOC 的 rootvg 使用 vSCSI 方式，相当于使用内置盘，便于排障 ；datavg 使用 NPIV，可以提高性能。以下是这种方式的优缺点，供参考，没有一种方式是绝对的好，需要综合实际环境去考虑：

优点	*由于 rootvg 使用 vscsi 的 MPIO，不使用额外的多路径软件，升级多路径软件比较容易*
	如果 VIOC 有 boot 的问题，容易排查故障
缺点	*VIOS 需要额外管理 VIOC rootvg 的磁盘*
	混合方式下，LPM 比只使用 NPIV 会比较复杂，需要提前把 VIOC rootvg 在 VIOS 上对应的磁盘在目标服务器上提前配置好

5.1.4 物理光纤卡的参数修改

对于 PowerVM 环境中的物理光纤卡，建议修改 fscsi 设备的 fc_err_recov 和 dyntrk 两个参数，命令如下：

```
$ chdev -dev fscsi0 -attr fc_err_recov=fast_fail dyntrk=yes
fscsi0 changed
```

在基于 PowerVM 的 PowerHA7.1 环境中，还建议将 fcs 设备的 tme 参数设置为 yes，命令如下：

```
$chdev -dev fcs0 -attr tme=yes
```

5.2 Virtual SCSI

VIOS 可以通过 vSCSI 把以下多种设备或者文件映射给 VIOC 使用：

- 物理卷
- 逻辑卷
- 文件比如 ISO 文件
- Shared storage pool 中的 Logical unit
- 光驱设备
- 磁带机设备

5.2.1 VIOS 上配置 virtual SCSI

对于 PowerVM 环境来讲，保持一个良好的记录习惯非常重要，对于后期的维护，变更，记录下物理设备，虚拟设备名称，对应关系等，建立一张维护表。对于 vSCSI 的 Slot 以及 Virtual Target Device(VTD)的命令，在 2.3.5 中已有介绍。

下面主要介绍如何在 VIOS 把一个 PV 通过 vSCSI 映射分 VIOC，把整个 PV 分配给 VIOC 有如下一些好处：

- 一个 PV 可以经由两个 VIOS 分配给 VIOC，为 VIOC 提供冗余 vSCSI 链路
- 一个 PV 可以同时分配给多个 VIOC 使用，比如作为集群的共享卷组
- 经由 SAN 提供给 VIOS 的 PV 再通过 VIOS 分配给 VIOC 支持 LPM 功能

在分配 PV 给 VIOC 之前，PV 的 reserve 属性需要修改为 no_reserve：

```
$ lsdev -dev hdisk12 -attr reserve_policy
value

single_path
$ chdev -dev hdisk12 -attr reserve_policy=no_reserve
hdisk12 changed
$ lsdev -dev hdisk12 -attr reserve_policy
value

no_reserve
```

PV 的 queue_depth 根据存储厂商提供的建议值修改。

5.2.2 VIOS 上配置 Virtual SCSI 设备给 Virtual I/O Client

相应的 vSCSI 适配器建好，PV 的属性修改完毕后，在 VIOS 上把 PV 通过 vSCSI 方式分配给 VIOC，使用如下命令：

```
$ mkvdev -vdev hdisk6 -vadapter vhost33 -dev aix02DrvgHd0
aix02DrvgHd0 Available
```

查看 vSCSI 的映射关系，使用如下命令：

```
$ lsmap -vadapter vhost33
SVSA            Physloc                                     Client Partition ID
--------------- ------------------------------------------- ------------------
vhost33         U8233.E8B.061AA6P-V1-C51                    0x00000005

VTD                   aix02rvgHd0

Status                Available
LUN                   0x8100000000000000
Backing device        hdisk6
Physloc               U5802.001.0086848-P1-C2-T1-W201600A0B829AC12-L3000000000000
Mirrored              false
```

如果需要删除 VSCSI 的配置，执行 rmvdev –vtd VTDname

5.2.3 VIOC 客户端配置 Virtual SCSI

由于 vSCSI 的 MPIO 只有 failover 功能，所以在 VIOC 上设置路径的优先级非常重要，比如在一个多 VIOC 的环境中，我们可以设置一部分 VIOC 的 vSCSI 路径优先通过第一个 VIOS，第二个 VIOS 作为备份通道，另一部分 VIOC 的 vSCSI 路径优先通过第二个 VIOS，第一个 VIOS 作为备份通道，这些操作需要在 VIOC 安装配置完毕后进行。

```
# lspath -l hdisk0
Enabled hdisk0 vscsi0
Enabled hdisk0 vscsi1
```

可以看到 VIOC 的 hdisk0 有两个 vSCSI 路径

```
# lspath -AHE -l hdisk0 -p vscsi0
attribute value description user_settable
priority  1     Priority    True
```

通过如上命令可以看到磁盘对应的 vSCSI 的优先级，在默认的情况下，优先级都为 1，1 表示最高，下面通过命令把这条路径的优先级改为 2，

```
# chpath -l hdisk0 -p vscsi0 -a priority=2
```

改完之后，表示 hdisk0 的 vSCSI 通道优先 vscsi1 对应的 VIOS，而 vscsi0 对应的 VIOS 作为备用通道，修改路径的优先级，不要重启 VIOC。

对于通过 vSCSI 映射给 VIOC 的磁盘，还建议修改如下属性的值,通过 chdev 命令修改，比如：

chdev -l hdisk0 -a reserve_policy=no_reserve -a algorithm=fail_over -a queue_depth=20 -a hcheck_mode=nonactive -a hcheck_interval=60 –P

chdev -l VSCSI0 -a VSCSI_path_to=30 -a VSCSI_err_recov=fast_fail –P

-P 参数标示只修改配置，需要重启成效。

如果修改的设备是 rootvg 使用，无法直接修改，需要加上参数-P，然后重启生效。

设备	***属性***	***建议值***
hdisk	algorithm	fail_over
	hcheck_interval	60 (需在 VIOS 和 VIOC 都要配置)
	hcheck_mode	nonactive
	queue_depth	与 VIOS 保持一致
	reserve_policy	no_reserve
vscsi	vscsi_err_recov	fast_fail (单 VIOS 的情况下，建议使用 delayed_fail)

	vscsi_path_to	30

5.3 NPIV 配置

本章节主要介绍 NPIV 的配置。对于 PowerVM 下的 SAN 存储环境，现在建议选择 NPIV 替代 VSCSI，优点如下：

- 存储上的 LUN 直接划分给 VIOC，不需要经过 VIOS，减少 VIOS 的管理
- VIOC 直接访问存储，多路径在 VIOC 上实现，可以实现 loadbalance
- VIOS 无需安装针对存储的多路径软件
- LPM 更简单
- 相比 VSCSI 可以提供更好的性能

NPIV 的一些特性以及考量点如下：

- 一个 VFC client 对应一个 VIOS 上的物理端口，避免单点
- 一个物理 FC 可以最多支持 64 个 VFC
- 一个物理 FC 对应的 VFC 最多支持 2048 个 WWPN
- 一个 POWER 服务器最多支持 32000 对 WWPN

5.3.1 VIOS 上配置 NPIV

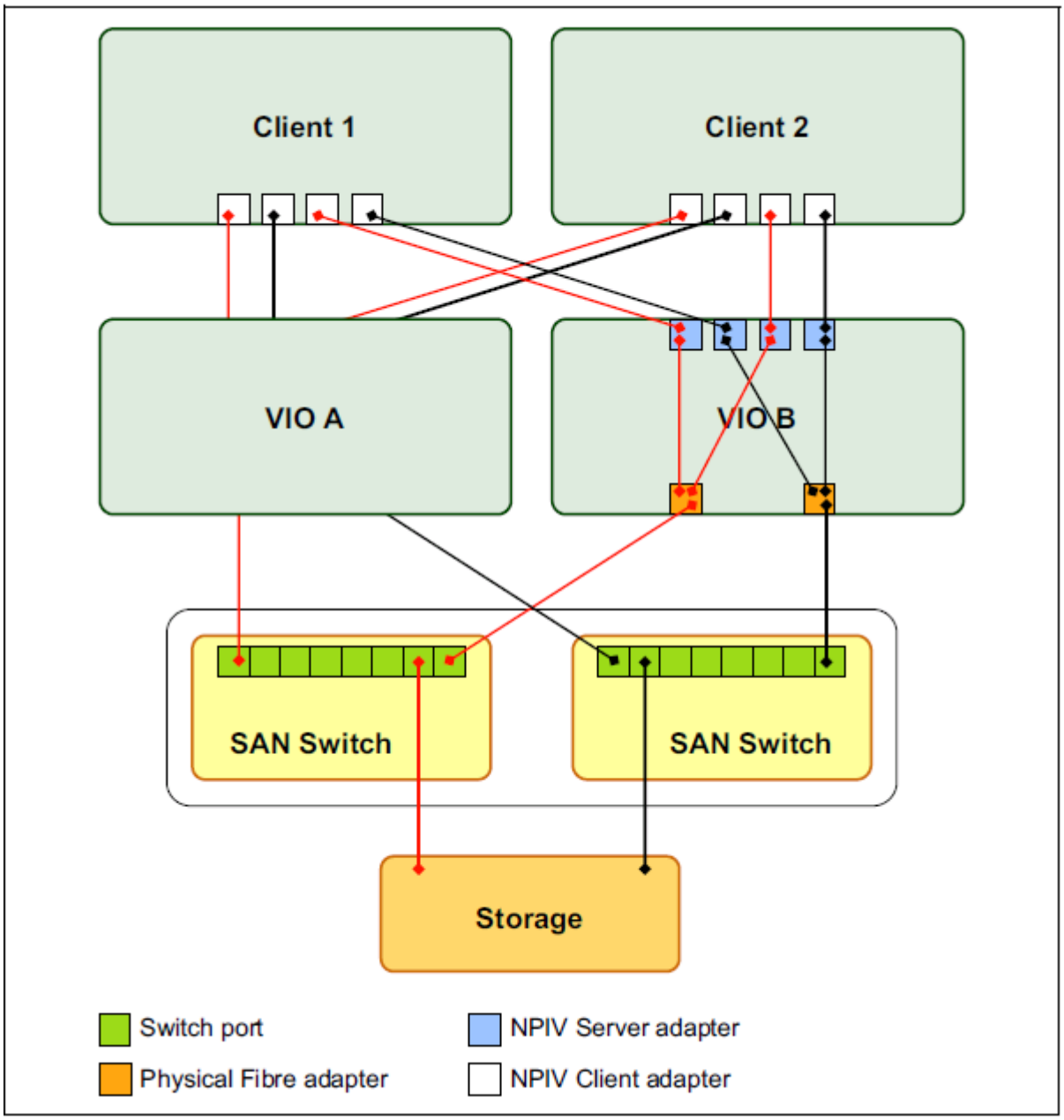
PowerVM 中支持 NPIV 需要以下前提条件：

- POWER6 以后的机器
- 8Gb 的光纤卡
- 支持 NPIV 的 SAN 交换机

- HMC V7.3.4 以后
- VIOS2.2 以上
- AIX5.3TL9 以上，AIX6.1TL2 以上
- SDD 1.7.2+PTF 1.7.2.2
- SDDPCM 2.2.0.0+PTF v2.2.0.6
- SDDPCM 2.4.0.0+PTF v2.4.0.1

对于 NPIV 配置来讲，也需要维护 VIOS 上 VFC 与物理 FC 的对应关系。VIOC 上的 VFC 会提供一对 wwpn，其中第一个 wwpn 在当前环境下使用，第二个 wwpn 在 LPM 的时候使用，当 VIOC 进行 LPM 操作后，迁移到另一台目标服务器上后，会使用第二个 wwpn，所以在 SAN 上划分 ZONE 的时候，需要把这对 wwpn 都划分在 ZONE 里，并且建议在 SAN 上划分 SOFT ZONE。

对于一个 SAN 环境，在各个层面提供冗余性是非常必要的，以下是一个典型的 SAN 环境：



使用 NPIV 的环境，建议如下：

- VIOS 建议多块网卡
- 建议使用双 VIOS
- 虚拟光纤卡对应不同的光纤卡，实现冗余
- 如果有备份通道，需要和数据通道使用不同的物理光纤卡
- 使用 NPIV 的话，对于 VIOC 来说相当于 SANBOOT，针对非 IBM 的存储，在 VIOC 安装操作系统的时候，建议 VIOC 先使用一条路径安装操作系统，操作系统安

装完毕后，安装存储厂商的多路径软件，安装完多路径软件后，再把 VFC 的其他路径加入。

5.3.2 VIOS 上配置 NPIV 设备给 Virtual I/O Client

首先修改 VIOS 上物理光纤卡两个参数 fc_err_recov 和 dyntrk，命令如下：

```
$ chdev -dev fscsi0 -attr fc_err_recov=fast_fail dyntrk=yes
fscsi0 changed
```

使用 lnports 查看可用的 NPIV 端口

```
$ lsnports
name             physloc                        fabric tports aports swwpns  awwpns
fcs0             U789D.001.DQD6L0C-P1-C1-T1          1     64     61   2048    2042
fcs2             U789D.001.DQD6K7W-P1-C1-T1          1     64     61   2048    2042
fcs4             U789D.001.DQD6L0M-P1-C1-T1          1     64     64   2048    2048
fcs6             U789D.001.DQD6L0M-P1-C3-T1          1     64     64   2048    2048
fcs8             U789D.001.DQD6K9W-P1-C1-T1          1     64     64   2048    2048
```

name 表示物理光纤卡在 VIOS 中的逻辑名

physloc 表示物理的位置号

fabric 为 1 表示支持 NPIV

tports 表示一个物理光纤卡可以支持 64 个虚拟光纤卡端口

aports 表示这块物理光纤卡支持剩余的虚拟光纤卡端口

swwpns 表示支持多少个 wwpn

awwpns 表示还剩余多少个 wwpn(其中 wwpn 用完即不可再重复利用)

然后执行 vfcmap 命令，建议 VIOS 上 virtual FC server adapter 与物理光纤卡的关系：

```
$vfcmap -vadapter vfchost0 -fcp fcs4
$vfcmap -vadapter vfchost1 -fcp fcs6
$vfcmap -vadapter vfchost2 -fcp fcs4
$vfcmap -vadapter vfchost3 -fcp fcs6
```

映射完成后，可以通过如下命令查看配置,lsmap –all –npiv:

```
Name          Physloc                          ClntID ClntName       ClntOS
------------- -------------------------------- ------ -------------- -------
vfchost0      U9179.MHB.1067E8R-V2-C30              4 IOMFBDB01_147  AIX

Status:LOGGED_IN
FC name:fcs4                   FC loc code:U78C0.001.DBJX588-P2-C1-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0           VFC client DRC:U9179.MHB.1067E8R-V4-C30

Name          Physloc                          ClntID ClntName       ClntOS
------------- -------------------------------- ------ -------------- -------
vfchost1      U9179.MHB.1067E8R-V2-C32              4 IOMFBDB01_147  AIX

Status:LOGGED_IN
FC name:fcs6                   FC loc code:U78C0.001.DBJX582-P2-C1-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs2           VFC client DRC:U9179.MHB.1067E8R-V4-C32

Name          Physloc                          ClntID ClntName       ClntOS
------------- -------------------------------- ------ -------------- -------
vfchost2      U9179.MHB.1067E8R-V2-C34              5 IOMEDDB01_149  AIX

Status:LOGGED_IN
FC name:fcs4                   FC loc code:U78C0.001.DBJX588-P2-C1-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0           VFC client DRC:U9179.MHB.1067E8R-V5-C34
```

其中 Status 为 LOGGED_IN 表示客户端的 VFC 已经在交换机上注册，如果客户端分区还未建立或者没有启动，此状态为 NOT_LOGGED_IN。

如果需要解除 VFC 与物理 FC 的关系，执行：

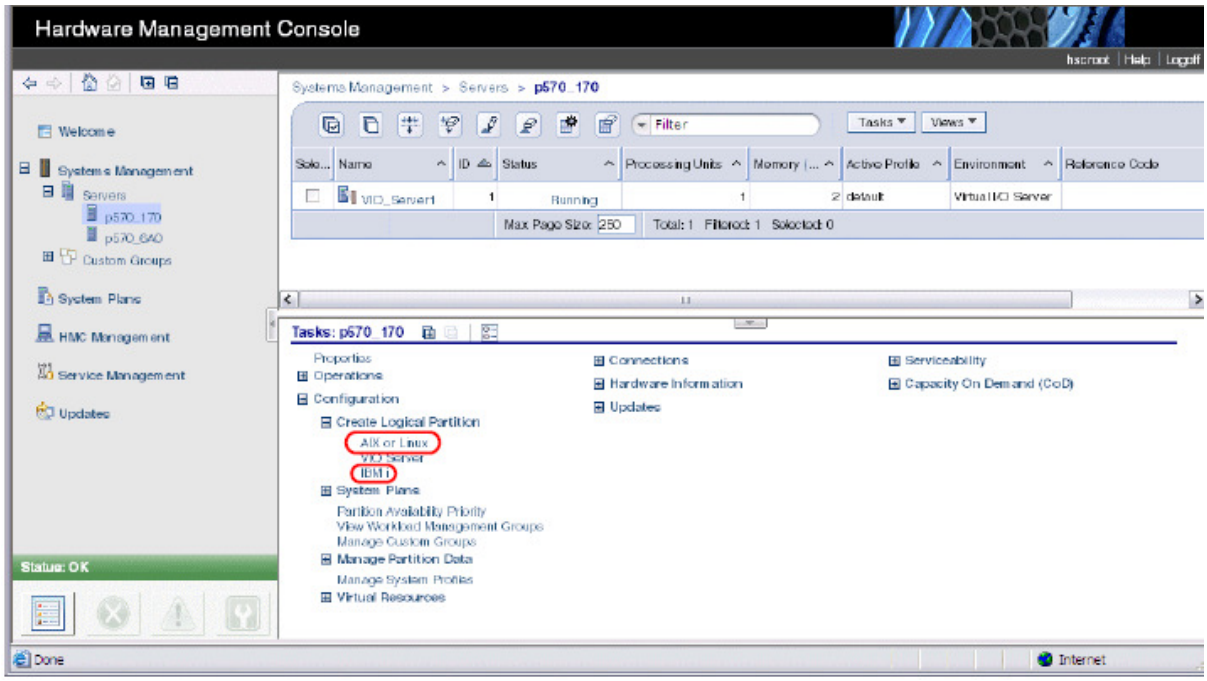
$ vfcmap -vadapter vfchost0 –fcp

6 客户端分区的安装配置

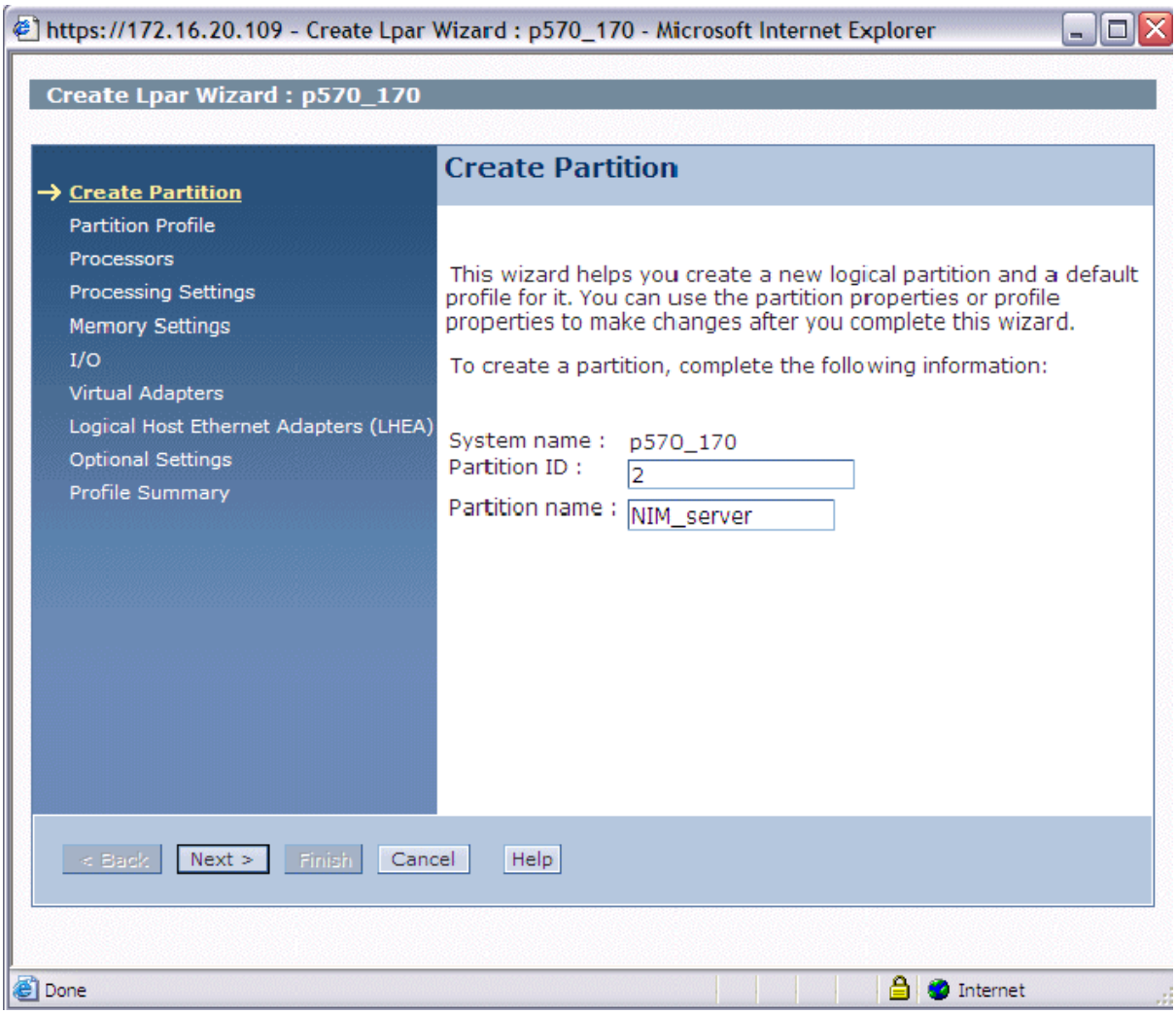
VIOS 设置完毕后，下面开始 VIOC 的设置

6.1 创建 VIO Client 分区

创建 VIOC 与创建 VIOS 类似，不同的是创建分区的时候选择 AIX 类型的分区



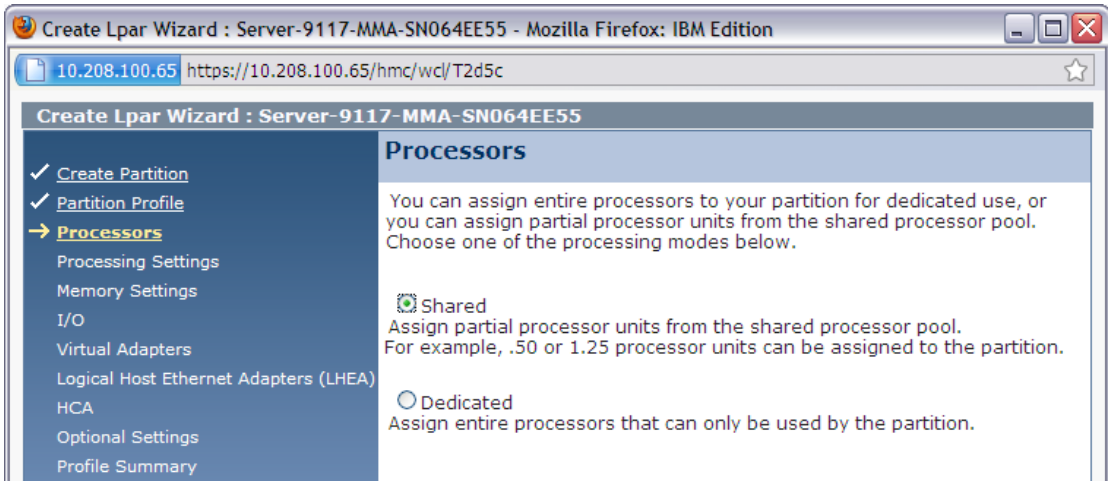
通过 HMC 上，然后选择“**Configuration -> Create Logical Partition -> VIO Server**”创建类型为 AIX 类型的逻辑分区



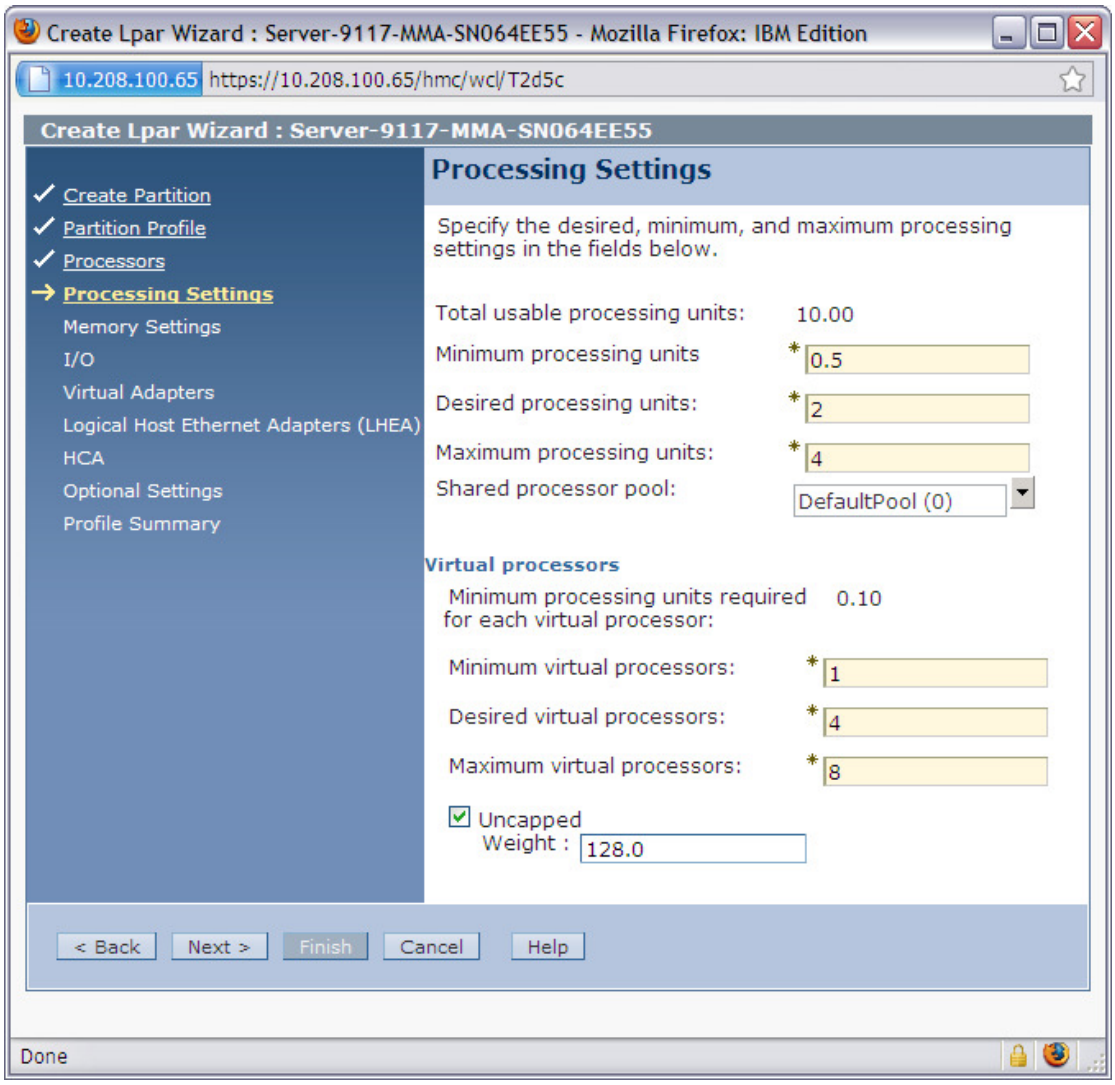
输入分区的分区 ID 和分区名称，点击下一步

输入概要文件的名称，点击下一步，进入 CPU 的设定

6.1.1 CPU 设定



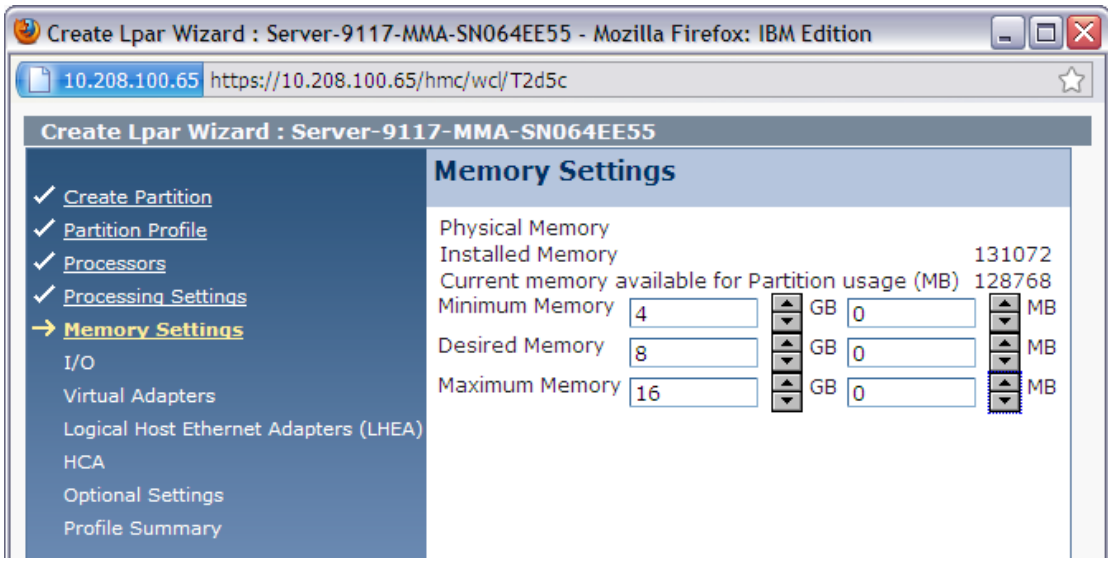
CPU 模式选择共享模式，



根据规划选择相应的值，点击下一步，进入内存设定

6.1.2 内存设定

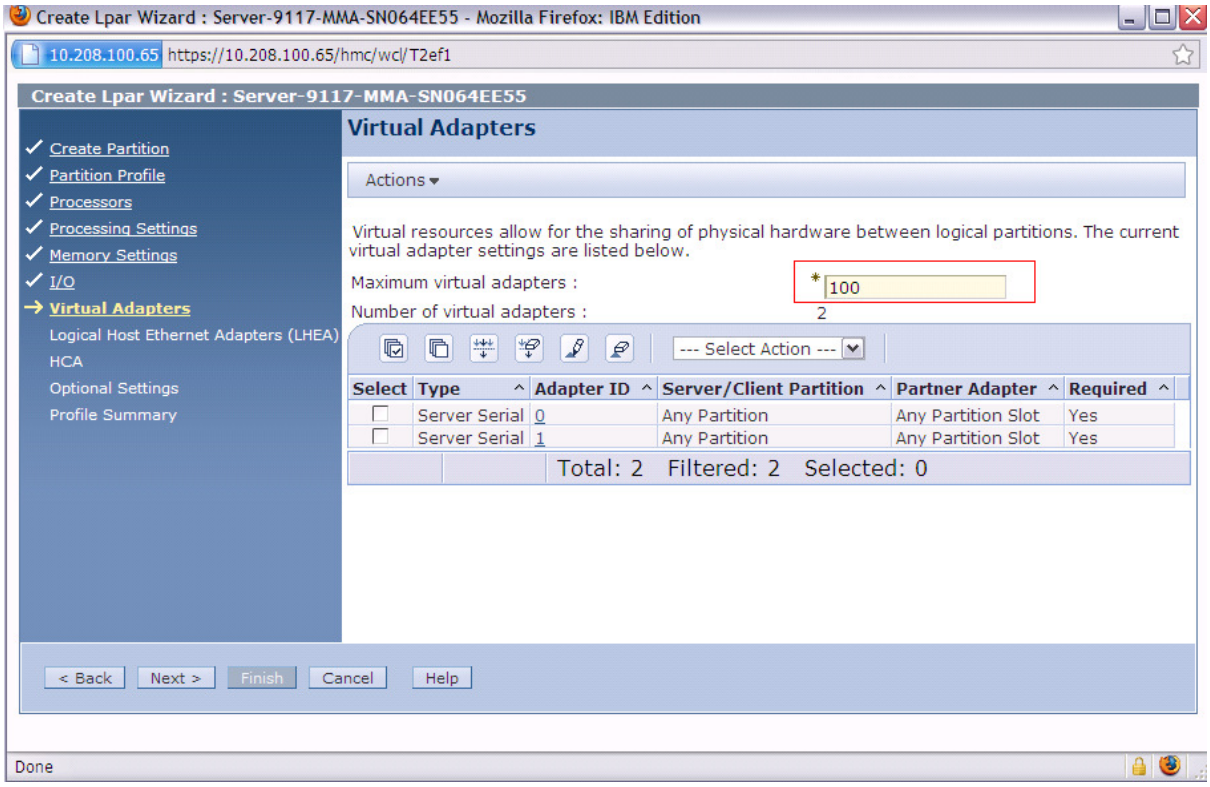
内存选择独占 Dedicated 模式



根据规划输入相应的值，进入下一步

6.1.3 虚拟 IO 适配器设定

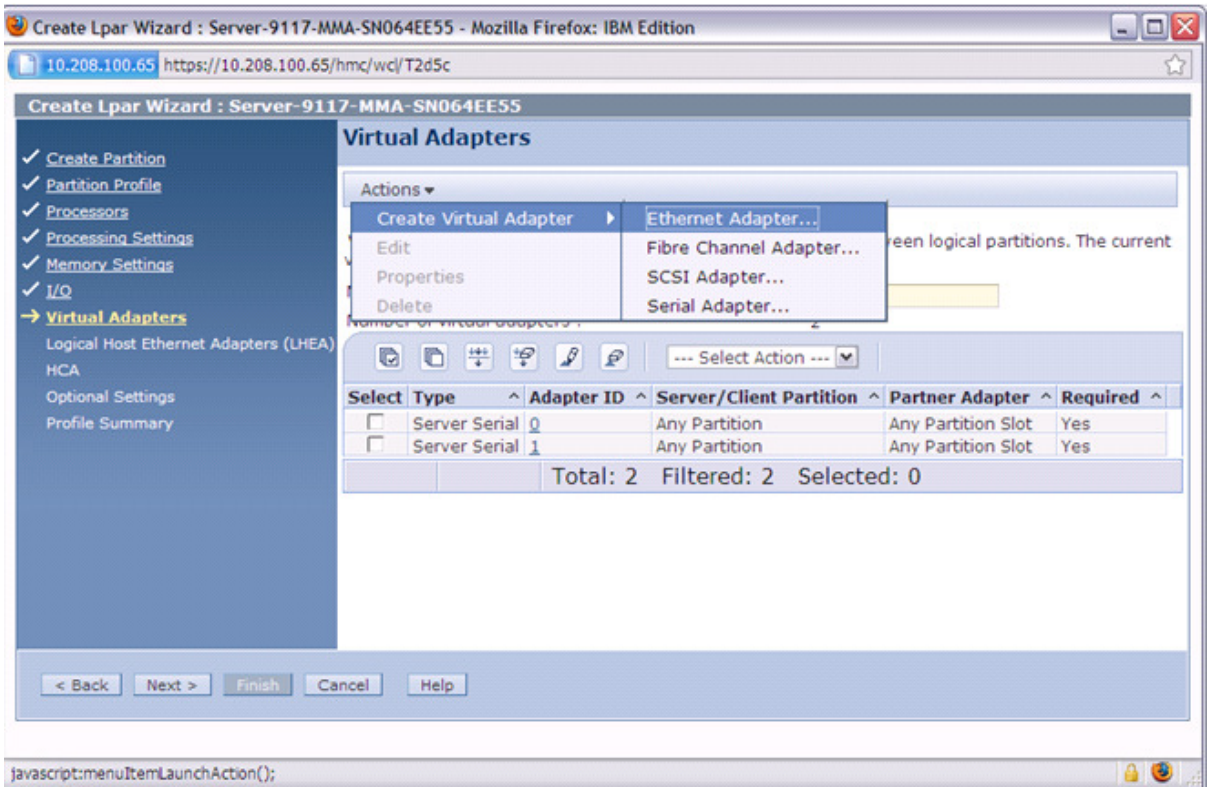
对于 VIOC 来讲，不分配任何的物理 IO 资源，直接进入虚拟 IO 适配器的设定

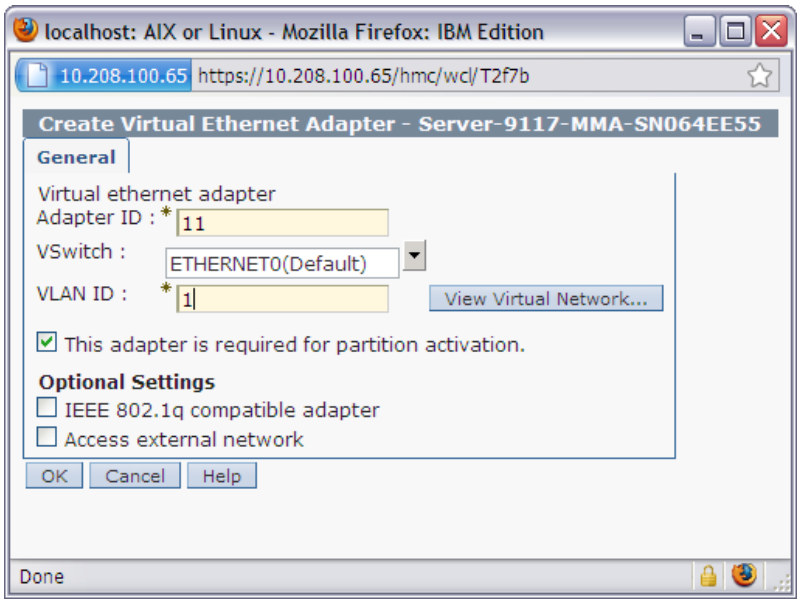


注意设置 Maximum virtual adapters 的值，以保证可以设置足够多的虚拟适配器

创建虚拟网卡：

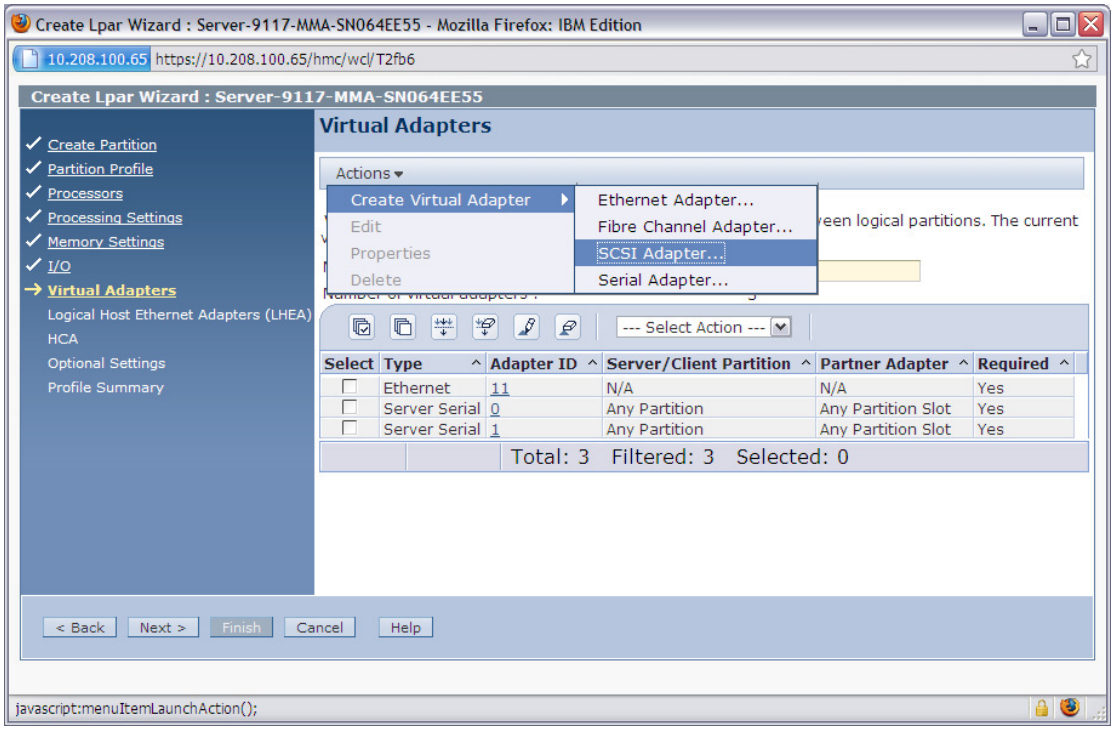
选择 Action→Create Virtual Adapter→Ethernet Adapter…

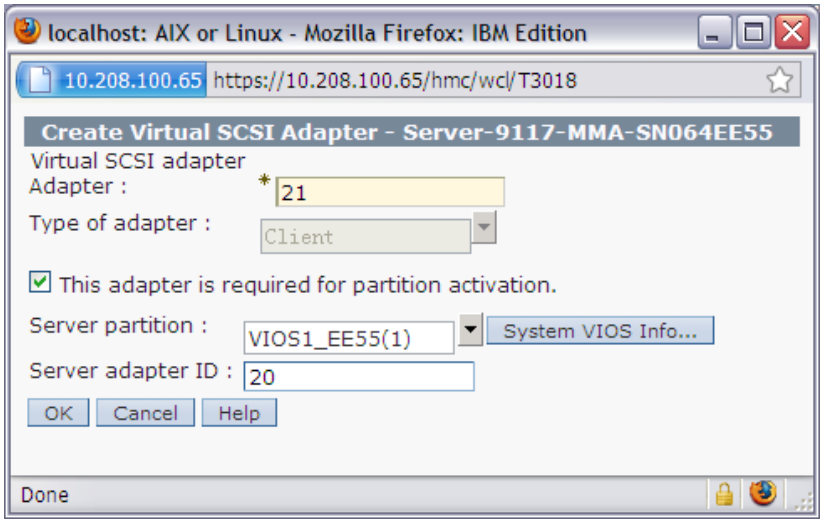




根据规划输入相应的 Slot ID，VLAN ID，选择默认的 VSwitch(ETHERNET0)，注意对于客户端分区的虚拟网卡来讲，不需要设置 Trunk，也不需要直接访问外部网络，所以 **Optional Settings** 一律为空。

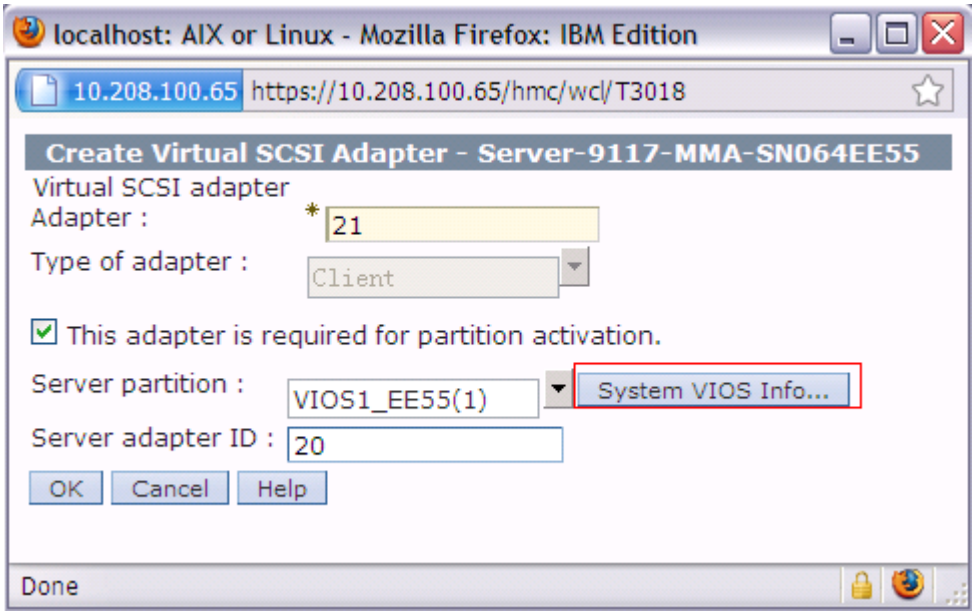
创建 VSCSI 卡：选择 Action→Create Virtual Adapter→SCSI Adapter…

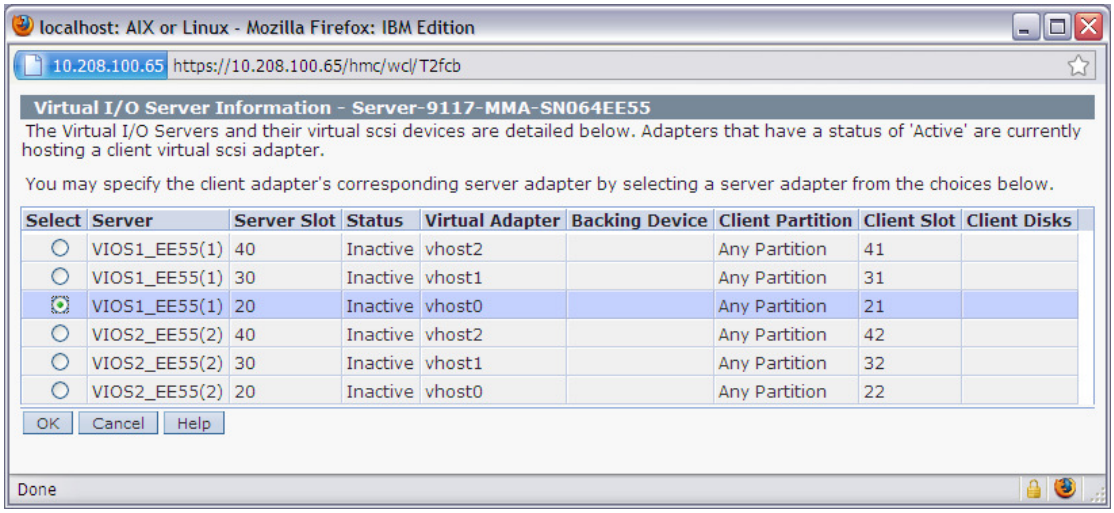




根据规划输入相应的 Slot ID，以及对应的 VIOS 上的 vSCSI Server adapter，点击 OK

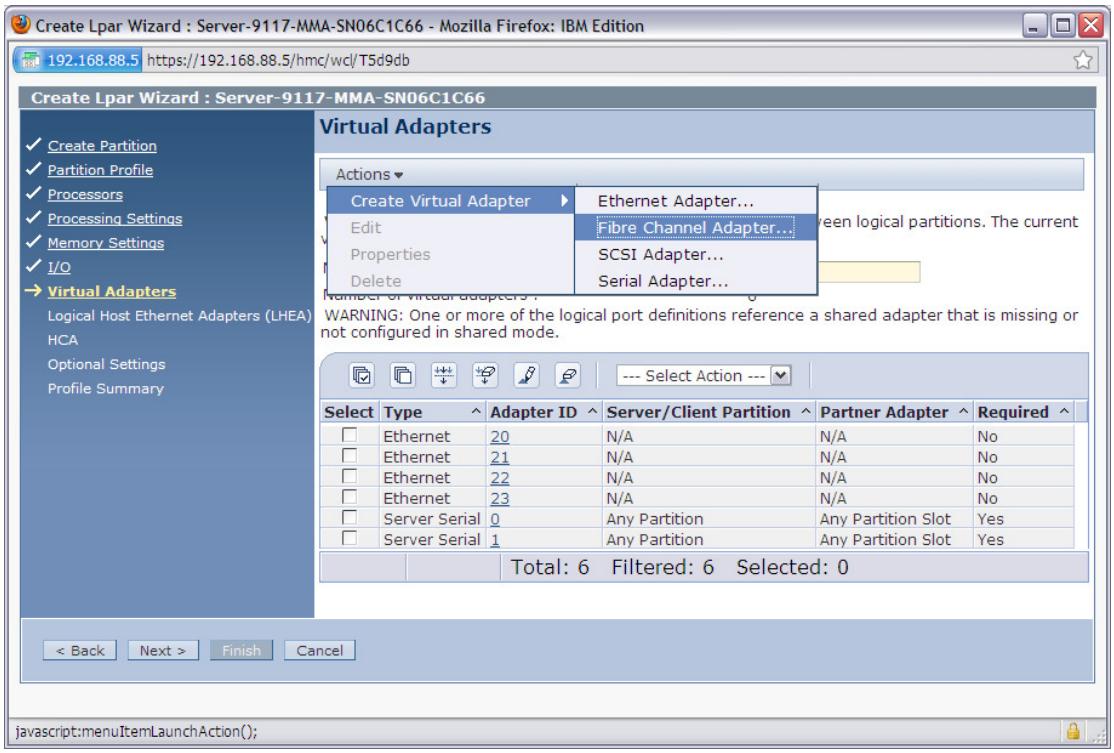
除了上面手工输入 VIOS 上 vSCSI Server adapter 以下；还有以下选择的办法，点击旁边的“System VIOS Info...”按钮，出现如下画面：

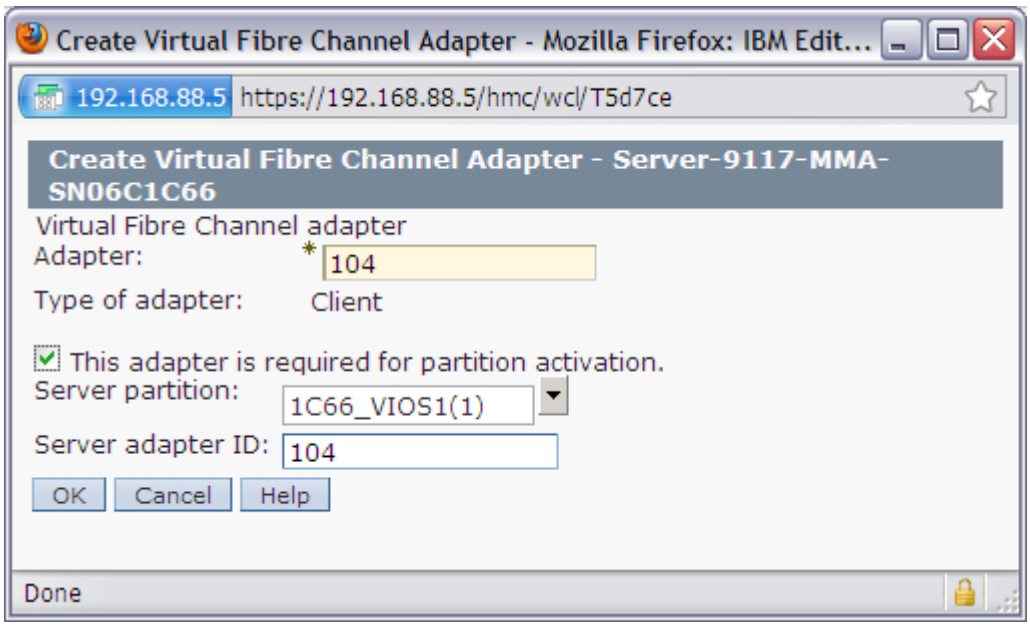




可以直接选择对应 VIOS 的 Virtual SCSI Server adapter。

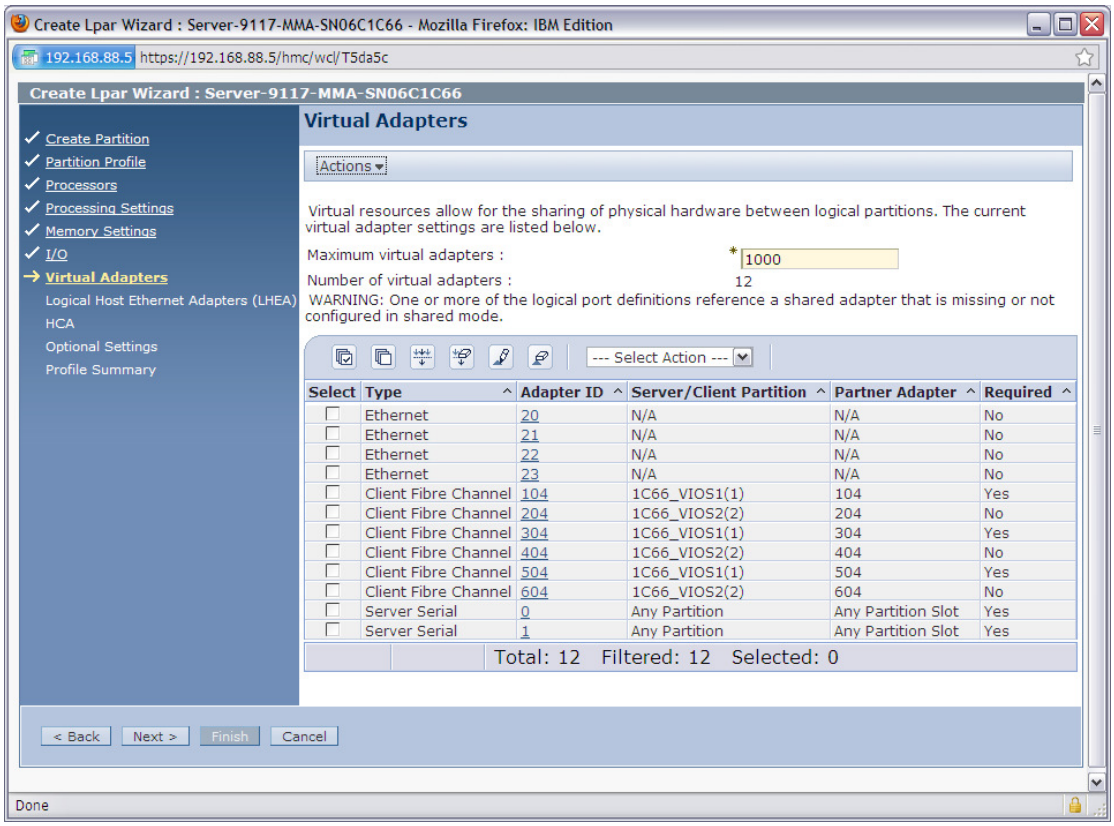
创建 VFC 卡：选择 Action→Create Virtual Adapter→Fibre Channel Adapter…





输入对应的 Slot ID，以及对应 VIOS 上的 Server adapter 的 ID，点击 OK。

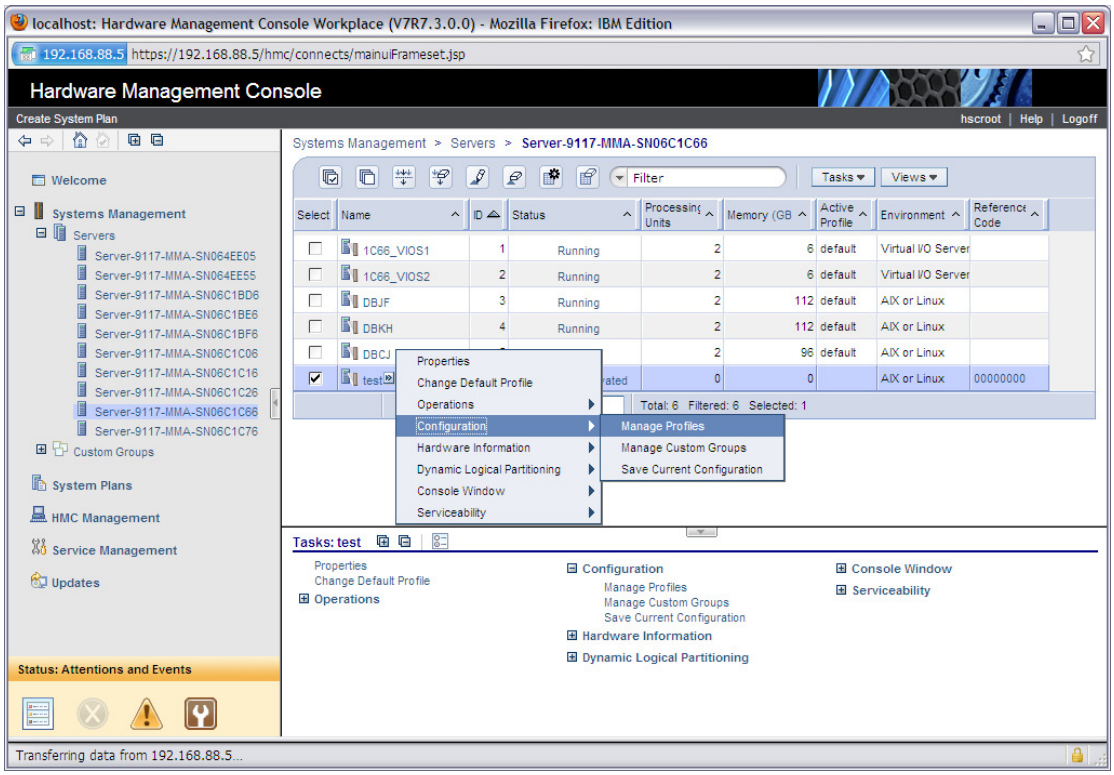
虚拟设备都创建完毕后，如下：



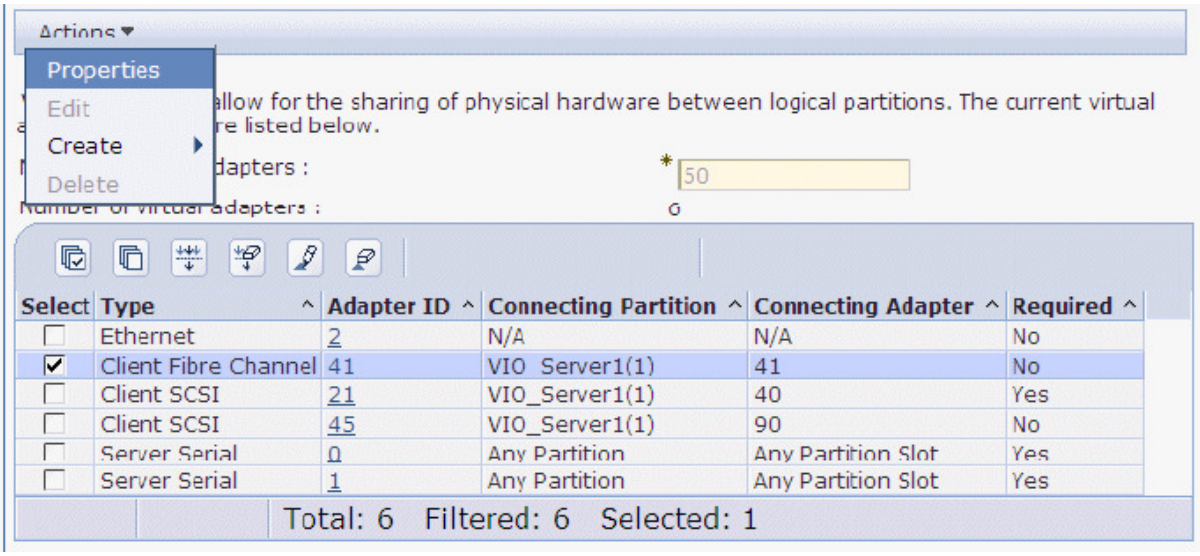
创建分区后面的选项都选择默认即可，直到分区创建结束。

6.1.3.1 NPIV 配置

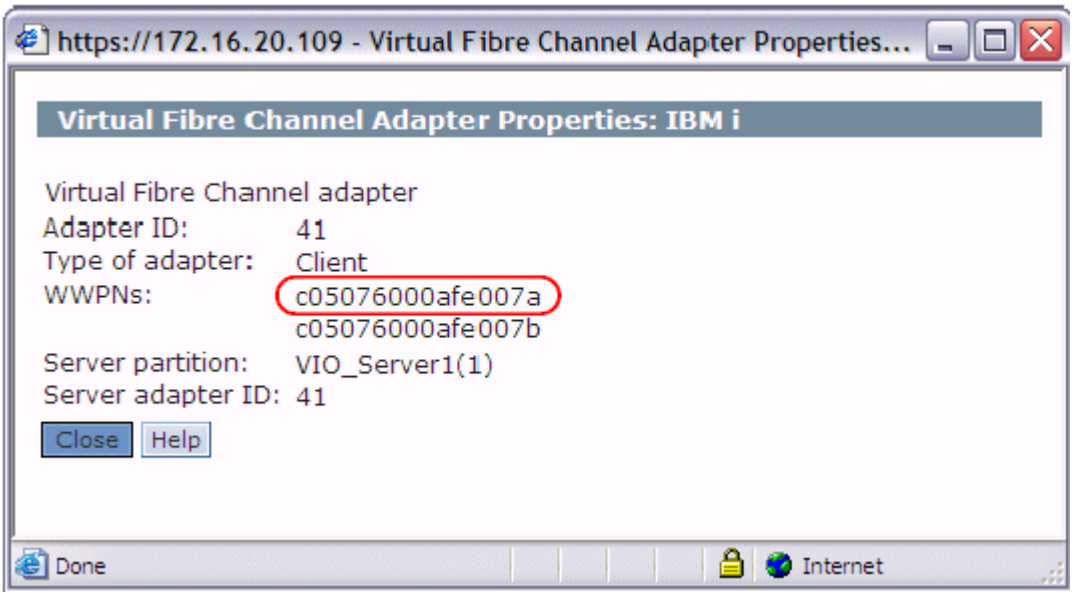
对于使用 NPIV 的 VIOC 来讲，我们需要知道 VIOC 上虚拟光纤卡的 wwpn，以便在 SAN 上创建 ZONE。可以通过如下办法查看：



选择一个 VIOC 分区，选择“Configuration->Manage Profiles”打开分区的概要文件



选择 VFC，点击属性查看



可以看到分区的一对 wwpn。

对于客户端的 VFC 卡的 wwpn，是 HMC 分配给这个虚拟卡的 WWPN 号，成对出现，这两个 WWPN 号如果要做 LPM，需要一起绑定到交换机（WWPN Zone）和存储。如果不需要，那么只要第一个就可以了。 要注意在对分区做 Dlpar 添加了 VFC 卡之后，不可以在用 manage profile 添加卡，这样会造成 Dlpar 产生的 VFC 卡和 Manage profile 产生的卡的 WWPN 号不一样，因次，需要使用 save current configuration 保存 DLPAR 之后的分区状态，或者直接使用 manage profile 修改之后再重启激活。

对应使用 NPIV SANBOOT 的客户端分区，如果是非 IBM 存储，建议在安装操作系统的时候使用单路径安装，安装完操作系统之后，安装存储厂商提供的多路径软件，再把其他的路径加入，不同的存储对于 SANBOOT 的处理也不同，可查询存储的产品手册。

6.2 客户端分区安装

AIX 分区的安装有多种办法，通过物理光驱、虚拟光驱或者通过 NIM，不再介绍。

AIX 安装完毕后，建议安装如下的 Bundle:

- Server
- App-Dev
- Alt_Disk_Install

- CDE
- openssh_client
- openssh_server

6.3 客户端分区升级

升级 AIX 需要登陆 IBM 官方网站下载升级补丁：

http://www-933.ibm.com/support/fixcentral/

每个升级包都有对应的 README，需要认真阅读，进行升级，不再介绍。

建议升级之前做系统备份。

6.4 客户端分区配置

客户端 AIX 安装、升级完毕后，根据不同应用的需要更改系统参数。

对于基于 VSCSI 方式的磁盘，建议通过 5.2.3 修改相应的参数值:

对每一个 hdisk，需要修改如下属性。注意在 vSCSI 模式下 queue_depth 和 hcheck_interval 与 VIOS 的对应值保持一致（见 VIOS 的参数配置）：

chdev -l hdisk0 -a reserve_policy=no_reserve -a algorithm=fail_over -a hcheck_mode=nonactive -a hcheck_interval=50 –P

修改 VSCSI Path 失效的超时时间，默认没启用

chdev -l vscsi0 -a vscsi_path_to=30 -a vscsi_err_recov=fast_fail -P

chdev -l vscsi1-a vscsi_path_to=30 -a vscsi_err_recov=fast_fail -P

对两条路径设定优先级：

chpath -l hdisk0 -p vscsi0 -a priority=1

chpath -l hdisk0 -p vscsi1 -a priority=2

其中 queue_depth 的值根据存储厂商的提供的建议值修改

对于基于 NPIV 方式的磁盘，只需修改磁盘的 reserver 属性为 no_reserver，queue_depth 的值根据存储厂商提供的建议值修改。

7 PowerVM 环境下的监控

本章节介绍 PowerVM 环境下 VIOS 中常见的监控手段。

7.1 PowerVM 环境下的监控

评估一个系统可以分为以下两种：

- 短期监控：作为基本的测试，容量调整以及排查故障
- 长期监控：监控性能的趋势、变化，容量管理

7.1.1 短期的监控手段

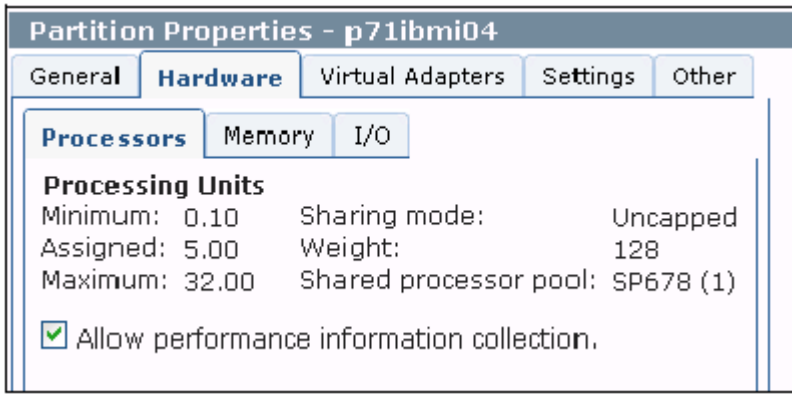
作为短期性能监控，常用的命令和工具如下：

命令或工具	说明
viostat	报告 VIOS 中 CPU、IO 的性能统计
netstat	报告网络的性能统计
vmstat	系统线程，虚拟内存，磁盘，CPU 的状态
topas	报告系统 CPU、内存、IO、网络的统计
seastat	报告 SEA 的性能统计
svmon	内存的使用统计
fcstat	光纤设备的使用统计

topas 可以提供大部分需要的数据，但是如果需要详细的数据，可以根据实际需要使用其他的几个工具。VIOS2.1 以后系统集成了 NMON 工具。

监控系统的 CPU 和共享处理器池

如果需要在分区中监控到共享处理器池的使用情况，首先需要通过 HMC 在 VIOS 分区上打开“Allow performance information collection.”选项



然后在此 VIOS 中通过 topas –fullscreen lpar 查看共享处理器池的使用情况，如下两个区域需重点查看：

Psize 共享处理器池中的物理 CPU 数量

app 共享处理器池中可用的物理 CPU 数量

同样在 AIX 中也可以通过 lparstat 命令来查看共享处理器池的使用情况。

磁盘 IO 使用监控：

在 VIOS 中短期监控磁盘 IO 的活动，使用 viostat 是个很好的办法，下面示例监控 vhost1 和 hdisk4 的活动状态，

```
$ viostat -adapter 1 1 | grep -p vhost1 | head -2 ; viostat -extdisk hdisk4 1 1
Vadapter:                 Kbps      tps     bkread    bkwrtn
vhost1                  88064.0    688.0      344.0     344.0
System configuration: lcpu=16 drives=12 paths=17 vdisks=41

hdisk4          xfer:  %tm_act      bps      tps      bread      bwrtn
                       100.0      90.2M   344.0        0.0       90.2M
                read:      rps  avgserv  minserv  maxserv   timeouts      fails
                           0.0      0.0      0.0      0.0          0          0
               write:      wps  avgserv  minserv  maxserv   timeouts      fails
                         344.0      7.1      5.3     13.4          0          0
               queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                           0.0      0.0      0.0      0.0        2.0        0.0
```

7.1.2 SEA 的监控

如果需要监控网络和 SEA 的状态，通常使用如下三个命令：entstat,netstat 和 seastat

netstat 命令：

netstat 提供了性能数据，网络信息如路由信息、网络数据，如下是 netstat 间隔为 1 秒的输出：

```
$ netstat 1
    input   (en11)     output           input   (Total)    output
 packets  errs  packets  errs colls  packets  errs  packets  errs colls
43424006     0    66342     0     0 43594413     0   236749     0     0
     142     0        3     0     0      142     0        3     0     0
     131     0        1     0     0      131     0        1     0     0
     145     0        1     0     0      145     0        1     0     0
     143     0        1     0     0      143     0        1     0     0
     139     0        1     0     0      139     0        1     0     0
     137     0        1     0     0      137     0        1     0     0
```

entstat 命令：

entstat 可以针对某块网卡提供统计信息，使用-all 参数显示所有的统计信息，在 VIOS 中使用 entstat 还可以查看 SEA 的状态和优先级，下面是 entstat 在一个双 VIOS 环境中其中一个 VIOS 上的示例：

```
$ entstat -all ent4 | grep Active
  Priority: 1  Active: True
```

可以看到 ent4 这个 SEA 的优先级是 1，当前是活动的状态

seastat 命令：

seastat 这个命令可以统计每个 VIOC 通过 SEA 的流量，使用之前首先需要打开 SEA 的 accounting 属性，命令如下：

```
$ lsdev -dev ent11 -attr | grep accounting
accounting    disabled Enable per-client accounting of network statistics True
$ chdev -dev ent11 -attr accounting=enabled
ent11 changed
```

当 accounting 属性打开后，SEA 会通过 VIOC 的 MAC 统计 VIOC 的流量，可以根据 IP 地址来统计，如下：

```
$ seastat -d ent9 -s ip=172.16.22.34
================================================================================
Advanced Statistics for SEA
Device Name: ent9
================================================================================
MAC: 22:5C:2B:95:67:04
----------------------
VLAN: None
VLAN Priority: None
IP: 172.16.22.34
Transmit Statistics:                 Receive Statistics:
--------------------                 -------------------
Packets: 29125                       Packets: 2946870
Bytes: 3745941                       Bytes: 184772445
================================================================================
```

7.1.3 长期的监控手段

如果需要观察 VIOS 长期的性能数据，使用趋势，就需要适用于长期监控的工具，下面逐一介绍：

topasrec 或者 topas_nmon:

topasrec 命令采用二进制格式来记录本地系统数据、跨分区数据（CEC 统计信息）和集群数据，在 VIOS 中默认的数据存放在/home/ios/perf/topas 目录，文件名的格式为 hostname.yymmdd,可以通过 ps 命令查看 toparec 是否在运行,topasrec 默认启动，在/etc/inittab 中定义，随系统启动而启动。

topas_nmon 的输出是 nmon 格式，可以通过如下方式打开，

cfgassist-

>Performance

->Topas

->Start New Recording->Start Persistent local recording

->nmon

默认文件存放在/home/ios/perf/topas 下

为了降低系统的性能，以上两种办法建议只使用一种，可以通过

cfgassist-

>Performance

->Topas

-> Stop Persistent recording 来关闭不需要的记录，同时注释/etc/inittab 的选项，以免系统启动后又启动。

VIOS 提供的供长期监控的 Agent：

VIOS 的监控也可以通过额外的产品来进行，VIOS 默认提供了如下的一些 Agent，可以供 ITM，IBM Systems Director 使用：

```
$ lssvc
ITM_premium
ITM_cec
TSM_base
ITUAM_base
TPC_data
TPC_fabric
DIRECTOR_agent
perfmgr
ipsec_tunnel
ILMT
```

HMC Utilization Data:

HMC utilization data 可以记录主机的 CPU 和内存使用情况，需要在每个服务器上启用 HMC 的”utilization data”功能，默认情况下没有启用，可以通过如下办法启用，选择一台服务器，选择”Operations->Utilization Data->Change Sampling Rate”，建议频率使用 5 分钟，开始记录数据；如果需要查看数据，选择” Operations->Utilization Data->View”查看

HMC 默认把数据记录在 HMC 的本地文件中，也可以通过 HMC 的命令列出某个服务器的数据，可以把这些数据导出，根据需要进行二次处理，如下：

```
hscroot@hmc9:~> lslparutil -m POWER7_1-SN061AA6P -r lpar --filter \
> "lpar_names=vios1a" -n 2 -F time,lpar_id,capped_cycles,\
> uncapped_cycles,entitled_cycles,time_cycles
06/08/2012
14:21:13,1,2426637226308,354316260938,58339456586721,58461063017452
06/08/2012
14:20:13,1,2425557392352,354159306792,58308640647690,58430247078431
```

VIOS Advisor:

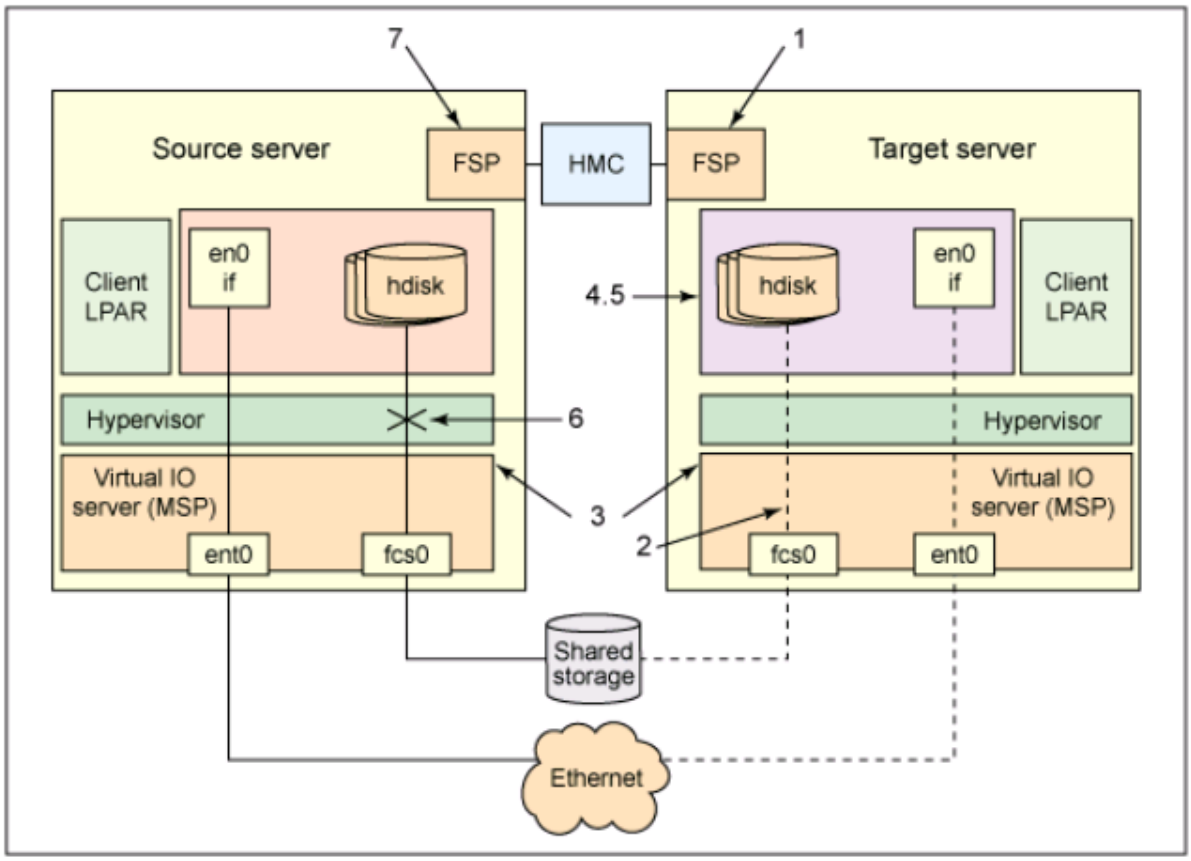
VIOS Advisor 可以收集 VIOS 一段时间内关键的系统数据，通过分析这些数据得出一个检查报告以及建议。建议 VIOS Advisor 在 VIOS 配完且有工作负载后，运行 24 小时的时间，可以得出 VIOS 的运行情况。

VIOS Advisor 是一个免费的工具 ，需要自行下载，介绍以及下载链接如下：

https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power+Systems/page/VIOS+Advisor

8 PowerVM 的高级特性

8.1 动态分区迁移 LPM(Live Partition Mobility)



动态分区迁移（Live Partition Mobility，以下简称 LPM）是 IBM 基于 POWER6 以后 技术提供的新特性，它特指将运行 AIX 或 Linux 操作系统的逻辑分区从一台物理系统迁移到另外一台完全不同的物理系统的过程。在这个过程中，操作系统和应用程序不受任何破坏，对外提供的服务也不受任何影响。

在实施 LPM 之前，需满足以下一些条件：

1. 迁移的分区没有使用任何本地的卡、磁盘、磁带机和光驱等物理 IO 设备
2. 两台服务器的 VIOS 必须能够同时访问共享存储
3. 两台服务器上分别要有 1 个 VIOS 被设置为 mover service partition，且这两个 VIOS

网络上要互相连通

4. 待迁移分区的 VLAN 在目标机器上必须存在，也可以访问外部网络
5. 待迁移分区的磁盘必须是共享存储上的磁盘，不可以是 VIOS 的逻辑卷或者本地磁盘
6. VIOS 分区必须可以做 DLPAR 操作
7. 如果是基于 vSCSI 的架构，待迁移分区的磁盘必须在目标服务器的两个 VIOS 上事先识别到，且 reserve 属性必须是 no_reserve;
8. 如果是基于 NPIV 的架构，待迁移分区的虚拟光纤卡的一对 wwpn 必须都划分在 ZONE 里，两个主机的 HBA 在数量和光纤通道的连接上是一致的，且磁盘的 reserve 属性必须是 no_reserve

8.2 内存共享 AMS(Active Memory Sharing)

共享内存是分配给共享内存池的物理内存，在多个逻辑分区之间共享该内存。共享内存池是已定义的一组物理内存块，它们由系统管理程序作为单个内存池进行管理。配置为使用共享内存的逻辑分区（以后称为共享内存分区）与其他共享内存分区共享池中的内存。

例如，创建具有 16GB 物理内存的共享内存池。然后创建三个逻辑分区，将它们配置为使用共享内存，并激活共享内存分区。每个共享内存分区可使用共享内存池中的 16GB。

分配给共享内存分区的内存量可大于共享内存池中的内存量。例如，可向共享内存分区 1 分配 12GB，向共享内存分区 2 分配 8GB，并向共享内存分区 3 分配 4GB。这些

共享内存分区总共使用 24GB 内存，但共享内存池只有 16GB 内存。在此情况下，内存配置被视为过量使用。

系统管理程序在共享内存分区的工作负载需求驱动下，通过不断执行下列任务来管理过量使用的内存配置：

* 根据需要，将来自共享内存池中的物理内存部分分配给共享内存分区；
* 根据需要，请求虚拟 I/O 服务器（VIOS）逻辑分区在共享内存池与调页空间设备之间读写数据。

在多个逻辑分区之间共享内存的能力称为 PowerVM Active Memory（活动内存）共享技术。PowerVM Active Memory（活动内存）共享技术是随 PowerVM 企业版提供的，您必须获取并输入 PowerVM 版激活码才能使用。

实现 Active Memory Sharing 的必要条件：

就像 Live Partition Mobility 一样，Active Memory Sharing 功能只能在 POWER6/POWER7 服务器中配置，同样 Active Memory Sharing 只能在 PowerVM 企业版中获得。如果客户目前使用的 PowerVM 易捷版或 PowerVM 标准版，那么首先需要升级到 PowerVM 企 业版，以便使用 Active Memory Sharing 功能。使用 Active Memory Sharing 功能的分区要求所有设备都是虚拟化的（通过 VIOS 获得），并定义此分区使用 Shared Processor 模式。

突出价值点：为内存根据应用需要在各分区间动态地、自动地切换提供了可能。

8.3 内存压缩 AME(Activate Memory Expansion)

Active Memory Expansion 是从 POWER7 开始支持的一项新的内存虚拟技术。它通过压缩 in-memory data 的方法，更加有效的利用内存。

目前支持 Active Memory Expansion 的要求是：

a) Power 710, 720, 730, 740, 750, 770, 780，激活了 AME 许可，运行在 POWER7 mode 下；

b) 服务器通过 HMC 管理: V7R7.1.0.0 或更新；

c) Firmware: 7.1；

d) AIX 6.1 TL4 SP2 或更新。

突出价值点：AME 通过内存数据压缩的方法，在已有内存的基础上，能够为一台服务器开辟更多的 LPAR，或者更加提高 LPAR 的性能。是更有效利用内存的很好的途径。

8.4 内存去重 AMD(Active Memory Deduplication)

AMD(Active Memory Deduplication)是 PowerVM 中的一项新技术。其基于 AMS（Active Memory Sharing)，通过对在 AMS 内存共享池内存储内容相同的内存空间进行去重，来达到优化内存使用的目的。Hypervisor 对在 AMS 共享池内的内 存页进行比对，如果发现相同内容的内存页，则通过映射来释放重复的内存空间，提高物理内存的使用效率。AMD 支持 AIX，i 和 Linux 分区。其对硬件的 微码要求为 740_042, 目前只有最新的 POWER7 C 类 Server 使用该级别的微码。

在 PowerVM 中，一个逻辑分区的内存并非指物理内存，而是逻辑内存。逻辑分区内存页与物理页之间的映射关系保存在一张表里，称之为 Hypervisor 逻辑内存表。当逻辑分区申请内存页的时候，PowerVM 将逻辑内存页地址转化为物理内存页地址。

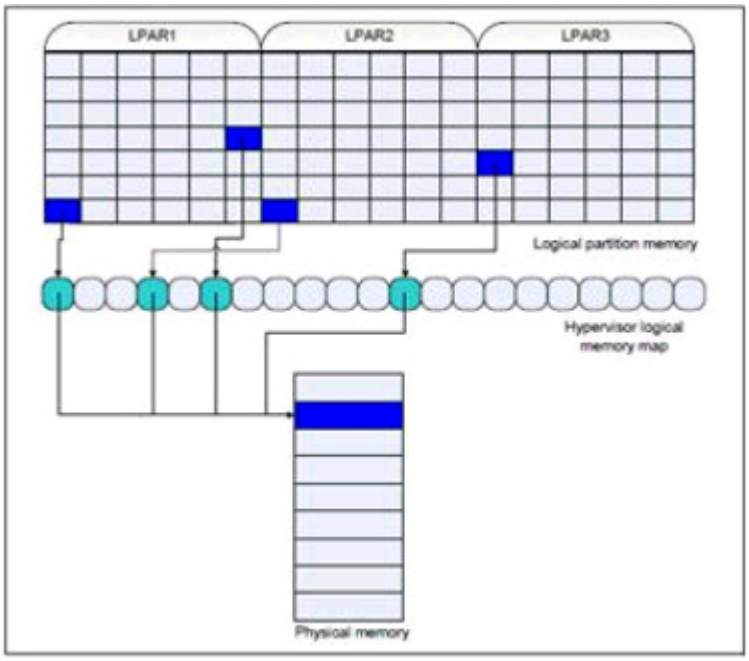
Hypervisor 逻辑内存表的优势是可以使一个逻辑分区内一段连续的逻辑内存块不连续的分布在物理内存中，这样，Hypervisor 在管理不同分区的内存时更加高效，分配更加灵活。

在正常的负载下，系统将数据存放在主内存中，并且在需要的时候，从主内存中读取数据。由于负载的特性，相同的数据，尤其是包含代码指令的数据可能被存放在内存中多个内存块中。

AMD 只有在配置了共享内存池时才能生效，一旦 AMD 功能被激活，连接到这个 AMS 的所有分区都将启用 AMD 功能，我们不能只对某几个分区启动这个功能。AMD 通过释放服务器的一个分区内或者分区之间重复内存页，来减少共享内存的过量使用，从而使主内存空间相同内存页面的数量最小化。为了优化内存利用率，AMD 避免在多个不同的物理内存空间之间做数据复制。为了实现这个目的，AMD 合并不同内存页面间相同数据，让数据只存在于一个内存页，然后释放其他具有相同数据的内存页。

例如，在 AMD 功能激活的情况下，当 Hypervisor 发现两个内存页具有相同的数据，重复页面释放的算法将会修改 Hypervisor 逻辑内存地址表，让逻辑分区的两个逻辑内存页都指向一个物理内存页，而另外一个内存页将会被释放。

下图清晰地展示了 AMD 的原理。在三个分区 lpar1,lpar2,lpar3 中均逻辑内存（蓝色所示），这些内存页中所包含相同的数据。通过 AMD 释放 AMS pool 上重复的内存页面并修改逻辑内存表的映射地址，三个逻辑指向 AMS pool 中同一块物理内存。这样，就避免不同的物理内存中出现重复的数据块。



附录 A：常用命令

在 VIOS 环境下，输入 help 即可列出所有的 VIOS 命令

常用的命令如下：

进入 Ksh：$ oem_setup_env

退出 Ksh：# exit

在$下扫描设备：$cfgdev

在$下配置：$ cfgassist

查看启动顺序：$ bootlist -mode normal -ls

查看物理卷：$ lspv

查看虚拟设备：$ lsdev –virtual

查看物理适配器：$ lsdev -type adapter

查看磁盘信息：$lsdev –type disk

查看虚拟以太网卡对应信息：$ lsmap -all -net

查看 SEA 状态：$entstat –all entX

查看虚拟 scsi 对应信息：$ lsmap -all | more

列出所有 vhost 映射：$ lsmap -all | grep vhost

查看镜像库：$ lsrep

磁盘映射给 vhost0：$ mkvdev -vdev hdisk4 -vadapter vhost0

删除磁盘映射：$ rmdev -dev vtscsi0

改变 SEA 的 HA 属性为 standby：$ chdev -dev ent4 -attr ha_mode=standby

查看 SEA 属性：$ lsdev -dev ent4 -attr

查看虚拟网卡优先级：$ entstat -all ent4 | grep Active

查看系统错误日志：$ errlog

查看哪些 HBA 卡是 NPIV 可用的：$ lsnports

NPIV 配置，把 vfchost9 和物理卡 fcs0 绑定到一起：$ vfcmap –vadapter vfchost9 –fcp fcs0

附录 B：参考资料

Redbook sg247590: <<IBM PowerVM Virtualization Managing and Monitoring>>

Redbook sg247940: <<IBM PowerVM Virtualization Introduction and Configuration>>